

# Accounting for Calibration Uncertainties in X-ray Analysis: Effective Areas in Spectral Fitting

Hyunsook Lee<sup>1</sup>, Vinay L. Kashyap<sup>1</sup>, David A. van Dyk<sup>2</sup>, Alanna Connors<sup>3</sup>,  
Jeremy J. Drake<sup>1</sup>, Rima Izem<sup>4</sup>, Xiao-Li Meng<sup>5</sup>, Shandong Min<sup>2</sup>,  
Taeyoung Park<sup>6</sup>, Pete Ratzlaff<sup>1</sup>, Aneta Siemiginowska<sup>1</sup>, and Andreas Zezas<sup>7,8</sup>

<sup>1</sup>Smithsonian Astrophysical Observatory, 60 Garden Street, Cambridge, MA 02138

`hlee@cfa.harvard.edu`

`vkashyap@cfa.harvard.edu`

`jdrake@cfa.harvard.edu`

`rpete@head.cfa.harvard.edu`

`asiemiginowska@cfa.harvard.edu`

<sup>2</sup> Department of Statistics, University of California, Irvine, CA 92697-1250

`dvd@ics.uci.edu`

`shandonm@uci.edu`

<sup>3</sup>Eureka Scientific, 2452 Delmer Street Suite 100, Oakland CA 94602-3017

`aconnors@eurekabayes.com`

<sup>4</sup>US Food and Drug Administration, Center for Drug Evaluation and Research,  
Division of Biometrics 4, 10903 New Hampshire Ave, Silver spring, MD 20903

`rima.izem@fda.hhs.gov`

<sup>5</sup>Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138

`meng@stat.harvard.edu`

<sup>6</sup>Department of Applied Statistics, Yonsei University, Seoul 120-749, South Korea

`taeyoung.t.park@gmail.com`

<sup>7</sup> IESL, Foundation for Research and Technology, 711 10, Heraklion, Crete, Greece

<sup>8</sup>Physics Department, University of Crete, P.O. Box 2208, 710 03, Heraklion, Crete, Greece

`azezas@physics.uoc.edu`

Received \_\_\_\_\_;    accepted \_\_\_\_\_

accepted for publication in ApJ

## ABSTRACT

While considerable advance has been made to account for statistical uncertainties in astronomical analyses, systematic instrumental uncertainties have been generally ignored. This can be crucial to a proper interpretation of analysis results because instrumental calibration uncertainty is a form of systematic uncertainty. Ignoring it can underestimate error bars and introduce bias into the fitted values of model parameters. Accounting for such uncertainties currently requires extensive case-specific simulations if using existing analysis packages. Here we present general statistical methods that incorporate calibration uncertainties into spectral analysis of high-energy data. We first present a method based on multiple imputation that can be applied with any fitting method, but is necessarily approximate. We then describe a more exact Bayesian approach that works in conjunction with a Markov chain Monte Carlo based fitting. We explore methods for improving computational efficiency, and in particular detail a method of summarizing calibration uncertainties with a principal component analysis of samples of plausible calibration files. This method is implemented using recently codified *Chandra* effective area uncertainties for low-resolution spectral analysis and is verified using both simulated and actual *Chandra* data. Our procedure for incorporating effective area uncertainty is easily generalized to other types of calibration uncertainties.

*Subject headings:* X-rays: general, methods: data analysis, methods: statistical, techniques: miscellaneous

## 1. Introduction

The importance of accounting for statistical errors is well established in astronomical analysis: a measurement is of little value without an estimate of its credible range. Various strategies have been developed to compute uncertainties resulting from the convolution of photon count data with *instrument calibration products* such as effective area curves, energy redistribution matrices, and point spread functions. A major component of these analyses is good knowledge of the instrument characteristics, described by the instrument calibration data. Without the transformation from measurement signals to physically interesting units afforded by the instrument calibration, the observational results cannot be understood in a meaningful way. However, even though it is well known that the measurements of the instrument’s properties (e.g., quantum efficiency of a CCD detector, point spread function of a telescope, etc.) have associated measurement uncertainties, the calibration of instruments is often taken on faith, with only nominal estimates used in data analysis, even when it is recognized that these uncertainties can cause large systematic errors in the inferred model parameters.<sup>1</sup> In many subfields (exceptions include: e.g. gravitational wave

---

<sup>1</sup>However in ground-based observations, it is customary to describe non-instrumental systematics as *calibration uncertainty*, especially time-variable and foreground effects, and incorporate them in the final uncertainties. These include: e.g. atmospheric absorption effects on photometry, flat-fielding, and astrometric calibration, as in Taris et al. 2011, Aguirre et al. 2011; calibrating brightness of distant objects in the presence of foreground dust (Conley et al 2011, Kim and Miquel 2006, Mandel et al. 2009). As well, uncertainties associated with the basic physics, such as e.g. specific stellar absorption lines (Thomas, Maraston, and Johansson 2010); or other model-mismatch uncertainties, such as intrinsic SN light-curve variations (Conley et al 2011, Kim and Miquel 2006, Mandel et al. 2009), can also be referred to as calibration uncertainties in the literature. In this paper, we specifically

astrophysics, VIRGO Collaboration 2010, LIGO Collaboration 2010 and references therein; CMB analyses, Mather et al. 1999, Rosset et al. 2010, Jarosik et al. 2011, and references therein; and extra-solar planet/planetary disk work, e.g. Butler et al. 1996, Maness et al. 2011, and references therein), instrument calibration uncertainty is often ignored entirely, or in some cases, it is assumed that the calibration error is uniform across an energy band or an image area. This can lead to erroneous interpretation of the data.

Calibration products are derived by comparing data from well-defined sources obtained in strictly controlled conditions with predictions, either in the lab or using a particularly well-understood astrophysical source. Parametrized models are fit to these data to derive best-fit parameters that are then used to derive the relevant calibration products. The errors on these best-fit values carry information on how accurately the calibration is known and could be used to account for calibration uncertainty in model fitting. Unfortunately, however, the errors on the fitted values are routinely discarded. Even beyond the errors in these fitted values, calibration products are subject to uncertainty stemming from differences between the idealized calibration experiments and the myriad of complex settings in which the products are used. Suspected systematic uncertainty cannot be fully understood until suitable data are acquired or cross-instrument comparisons are made (David et al. 2007). Prospectively, this source of uncertainty is difficult to quantify but is encompassed to a certain extent in the experience of the calibration scientists. Different mechanisms have been proposed to quantify this type of uncertainty, ranging from adopting ad hoc distributions such as truncated Gaussian (Drake et al. 2006) to uniform deviations over a specified range. As long as it can be characterized even loosely, statistical theory provides a mechanism by which this information can be included to better estimate the

---

concentrate on instrumental calibration uncertainties, although the formalisms introduced could in principle handle other kinds of systematic errors.

errors in the final analysis.

Users and instrument builders agree that incorporating calibration uncertainty is important (see Davis 2001; Drake et al. 2006; Grimm et al. 2009). For example, Drake et al. (2006) demonstrated that error bars on spectral model parameters are underestimated by as much as a factor of 5 (see their Figure 5) for high counts data when calibration uncertainty is ignored ( $\gg 10^3$  counts for typical CCD resolution spectra). Such underestimations can lead to incorrect interpretations of the analysis results. Despite this, calibration uncertainties are rarely incorporated because only a few ad hoc techniques exist and no robust principled method is available. In short, there is no common language or standard procedure to account for calibration uncertainty.

Historically, at the International Congress of Radiology and Electricity held in Brussels in September 1910, MMe. Curie was asked to prepare the first standard based on high energy photon emission (X-/ $\gamma$ -ray): 21.99 milligrams of pure radium chloride in a sealed glass tube, equivalent to  $1.67 \times 10^{-2}$  Curies of radioactive radium (e.g., Brown 1997 pg 9ff and references therein). The problem then became: how to measure other samples, in reference to this standard? Although the sample preparation was done by very accurate chemistry techniques, the tricky part was designing and building the instrument to quantify the high-energy photon emission. At the next International Committee meeting (1912, Paris) calibrating the standard was done by specialized electroscopes balancing the ‘ionization current’ from two sources. This instrument was deemed to have an uncertainty of one part in 400 (Rutherford and Chadwick 1911). The original paper also describes a method for calibrating the detector. Although these measurements were quite carefully done, and complex for their time, the result was a single value (the intensity) and had a single number quantifying its error ( $\frac{1}{400}$ ; Rutherford and Chadwick 1911). In this case, the effect of this original unavoidable measurement error on one’s final measurement of a source intensity (in

Curies) is straightforward to propagate, such as by the delta-method.

Nowadays, meetings about absolute standards and measuring instruments are much more complex, incorporating multiple kinds of measurements for a single standard (e.g. CODATA; Mohr, Taylor, and Newell 2008). As well, in the general literature, one finds increasingly complex methods dealing with e.g. multivariate data and calibration (Sundberg 1999, Osbourne 1991), and even methods for ‘traceability’ back to known standards (Cox and Harris 2006). These approaches formulate their complexities in terms of cross-correlations of parameters. This methodology has also been successfully used in modern astrophysics, such as in combining optical observations of supernovae for cosmological purposes (e.g. Kim and Miquel 2006). Initially, J. Drake and other co-authors did try formulating the dependencies and anticorrelations of the final calibration product uncertainties in terms of correlation coefficients. However, after considerable exploration, they found this approach unable to capture the complexities of spacecraft calibration, especially at high energies. First, each part of a modern instrument such as the Chandra observatory is measured at multiple energies and multiple positions, as well as calibrating the whole system on the ground. Second, interestingly, the instrument is modeled by a complex physics-based computer code. The original calibration measurements are not used directly, but are benchmarks for the physical systems modeled therein. High energy astrophysics brings a third difficulty: the previous papers assumed a Gauss-Normal distribution for the calibration-product uncertainties; this certainly does not hold for most real instruments in the high energy regime. Hence, expanding beyond Drake et al. (2006), in this paper, we describe how to ‘short-circuit’ tracing back to the original calibration uncertainties by using the entire instrument-modeling code as part of statistical computing techniques. We see this in the context of the movement towards “uncertainty quantification” (UQ) of large computer codes (see, e.g., Christie et al. 2005).

Until recently, the best available general strategy in high-energy astrophysics was to compute the root-mean-square of the measurement errors and the calibration errors and then to fit the source model using the resulting error sum (see Bevington and Robinson 1992). Unfortunately, the use and interpretation of the standard deviation relies on Gaussian errors, that the calibration errors are uncorrelated, and that the uncertainty on the calibration products can be uniquely translated to an uncertainty in each bin in data space. None of these assumptions are warranted. Furthermore, this method, equivalent to artificially inflating the statistical uncertainty on the data, will lead to biased fits, error bars without proper coverage, and incorrect estimates of goodness of fit. Individual groups have also tried various instrument-specific methods. These range from bootstrapping (Simpson and Mayer-Hasselwander 1986) to raising and lowering response “wings” by hand (Forrest 1988, Forrest Vestrand and McConnell 1997), and in one case, analytical marginalization over a particular kind of instrumental uncertainty (Bridle et al. 2002). In general and in important cross-instrument comparisons, however, all but the crudest methods (e.g., multiplying each instrument’s total effective area by a fitted “uncertainty factor” as in Hanlon et al. 1995, Schmelz et al. 2009) are very difficult to handle.

Methods for handling systematic errors exist in other fields such as particle physics (Heinrich and Lyons 2007 and references therein) and observational cosmology (Bridle et al. 2002). In their review of systematic errors, Heinrich and Lyons (2007) advocate parameterizing the systematics into statistical models and marginalizing over the nuisance parameters of the systematics. They described various statistical strategies to incorporate systematic errors which range from simple brute force  $\chi^2$  fitting to fully Bayesian hierarchical modeling. Unfortunately these analytical methods rely on Gaussian model assumption that are inappropriate for high energy astrophysics and are also highly case specific.

Accounting for calibration uncertainty is further complicated by complex and large



scale correlation in the calibration products. The value of the calibration product at one point can depend strongly on far away values and even data collected using a different instrument. For example, the *Chandra* Low Energy Transmission Grating Spectrometer (LETGS) + High Resolution Camera - Spectroscopic readout (HRC-S) effective area is calibrated using the power-law source PKS 2155-304. Because the high-order contributions to the spectrum cannot be disentangled, the index of the power-law depends strongly on an analysis of the same source with data obtained contemporaneously with the High Energy Transmission Grating Spectrometer (HETGS) + ACIS-S. Thus, changes in the HETGS+ACIS-S effective area will affect the longer-wavelength LETGS+HRC-S effective area. The complex correlations can result in a diverse set of plausible effective area curves. The choice among these curves can strongly affect the final best fit in day-to-day analyses. The nominally better strategy of folding the calibration uncertainty through to the final statistical errors on fitted model parameters is unfortunately unfeasible: the complex correlations make it difficult to quantify the affect on the final analysis of uncertainty in the calibration product.

Drake et al. (2006) proposed a strategy that accounts for these correlations by generating synthetic datasets from a nominal effective area and then fitting a model separately using each of a number of instance of a simulated effective area and then estimating the effect of the calibration error via the variance in the resulting fitted model parameters. This procedure can be implemented using standard software packages such as *XSPEC* (Arnaud 1996) and *Sherpa* (Freeman et al. 2001, Refsdal et al. 2009) and demonstrates the importance of including calibration errors in data analysis. However, in practice there are some difficulties in implementing it with real data where the true parameters are not known *a priori*. The ad hoc nature of the bootstrapping-type procedure means its statistical properties are not well understood, requiring the sampling distributions to be calibrated on a case-by-case basis. That is, the procedure requires verification

whenever different models are considered or different parts of the parameter space are explored. The large number of fits required also imposes a heavy computational cost. Most importantly, it requires numerous simulated calibration products that must be supplied to end users either directly through a comprehensive database or through instrument specific software for generating them. In general, both these strategies impose a heavy burden on calibration or analysis software maintainers.

The primary objective of this article is to propose well-defined and general methods to incorporate complex calibration uncertainty into spectral analysis in a manner that can be replicated in general practice without precise calibration expertise. Although we develop a general framework for incorporating calibration uncertainty, we limit our detailed discussion to accounting for uncertainty in the effective area for *Chandra*/ACIS-S in spectral analysis. We propose a Bayesian framework, where knowledge of calibration uncertainties is quantified through a prior probability. In this way, information quantified by calibration scientists can be incorporated into a coherent statistical analysis. Operationally, this involves fitting a highly-structured statistical model that does not assume the calibration products are known fixed quantities, but rather incorporates their uncertainty through a prior distribution. We describe two statistical strategies below for incorporating this uncertainty into the final fit. Multiple imputation fits the model several times using standard fitting routines, but with a different value of the calibration product used in each fit. Alternatively, using an iterative Markov chain Monte Carlo (MCMC) sampler allows us to incorporate calibration uncertainty directly into the fitting routine by updating the calibration products at each iteration. In either case, we advocate updating the calibration products based solely on information provided by calibration scientists and not on the data being analyzed (i.e., not updating products given the data being analyzed; see also discussion about computational feasibility in §6.1). This strategy leads to simplified computation and reliance on the expertise of the calibration scientists rather than on the idiosyncratic features of the data.

We adopt the strategy of Drake et al. (2006) to quantify calibration uncertainty using an ensemble of simulated calibration products, that we call the *calibration sample*. We use Principal Component Analysis (PCA) to simplify this representation. A glossary of the terms and symbols that we use is given in Table 1.

In §2 we describe the calibration sample and illustrate the importance of properly accounting for calibration uncertainty in spectral analysis. Our basic methodology is outlined in §3, where we describe how the calibration sampler can be used to generate the replicates necessary for multiple imputation or can be incorporated into an MCMC fitting algorithm. We also show how PCA can provide a concise summary of the complex correlations of the calibration uncertainty. Specific algorithms and strategies for implementing this general framework for spectral analysis appear in §4. Our proposed methods are illustrated with a simulation study and an analysis of 15 radio loud quasars (Siemiginowska et al. 2008) in §5. In §6 we discuss future directions and a general framework for handling calibration uncertainties from astrophysical observations with similar form as our  $\gamma$ -ray examples. We summarize the work in §7.

## 2. The Calibration Sample and the Effect of Calibration Uncertainty

To coherently and conveniently incorporate calibration uncertainty into spectral fitting, we follow the suggestion of Drake et al. (2006) to represent it using a randomly generated set of calibration products that we call the *calibration sample*. In this section we begin by describing this calibration sample, and how it can be used to represent the inherent systematic uncertainty. The methods that we discuss in this and the following sections are quite general and in principle can be applied to account for systematic uncertainty in any calibration product. For clarity, we illustrate their application to instrument effective areas.

## 2.1. Representing Uncertainty

We begin with a simple model of telescope response that assumes position and time invariance. In particular, suppose the response of a detector to an incident photon spectrum  $S(E; \theta)$ ,

$$\mathcal{M}(E^*; \theta) = \sum_E S(E; \theta) A(E) P(E) R(E^*; E), \quad (1)$$

where  $E^*$  represents the detector channel at which a photon of energy  $E$  is recorded,  $\theta$  represents the parameters of the source model, and  $A$ ,  $P$ , and  $R$  are the effective area, point spread function, and energy redistribution matrix of the detector, respectively. We aim to develop methods to estimate  $\theta$  and compute error bars that properly account for uncertainty in  $A$ . Of course  $P$  and  $R$  are also subject to uncertainty and in §6.2 we discuss extensions of the methods described here to handle more general sources of calibration uncertainty.

As an illustration, we consider observations obtained using the spectroscopic array of the *Chandra* AXAF CCD Imaging Spectrometer detector (ACIS-S). According to Drake et al. (2006), it is possible to generate a calibration sample of effective area curves for this instrument by explicitly including uncertainties in each of its subsystems (UV/Ion shield transmittance, CCD Quantum Efficiency, and the telescope mirror reflectivity). The result is a set of simulations of the effective area curves. These encompass the range of its uncertainty, with more of the simulated curves similar to its most likely value, and fewer curves that represent possible but less likely values. In principle, some may be more likely than others, in which case weights that indicate the relative likelihood are required. In this article, we assume that all of the simulations in the set are equally likely, that is the simulations are representative of calibration uncertainty. The set of  $L$  simulations is the *calibration sample* and denoted  $\mathcal{A} = \{A_1, \dots, A_L\}$ , where  $A_l$  is one of the simulated effective area curves.

The complicated structure in the uncertainty for the true effective area is illustrated in Figure 1 using the calibration sample of size  $L = 1000$  generated by Drake et al. (2006). A selection of six of the  $A_l$  from  $\mathcal{A}$  are plotted as colored dashed lines and compared with the default effective area,  $A_0$  that is plotted as a solid black line. The second panel plots the differences,  $A_l - A_0$  for the same selection. The light gray area represents the full range of  $\mathcal{A}$  and the dark gray area represents intervals that contain 68.3% of the  $A_l$  at each energy. The complexity of the uncertainty of  $A$  is evident. We use the calibration sample illustrated in Figure 1 as the representative example throughout this article.

## 2.2. The Effect of the Uncertainty

We discuss here the effect of the uncertainty represented by the calibration sample on fitted spectral parameters and their error bars. We employ simulated spectra representing a broad range in parameter values. In particular, we simulated four data sets of an absorbed power-law source with three parameters (power-law index  $\Gamma$ , absorption column density  $N_H$ , and normalization) using the `fakeit` routine in `XSPECv12`. The data sets were all simulated without background contamination using the `XSPEC` model `wabs*powerlaw`, nominal default effective area  $A_0$  from the calibration sample of Drake et al. (2006), and a default RMF for ACIS-S. The power law parameter ( $\Gamma$ ), column density ( $N_H$ ), and nominal counts for the four simulations (see also Table 2) were

**SIMULATION 1:**  $\Gamma = 2$ ,  $N_H = 10^{23}\text{cm}^{-2}$ , and  $10^5$  counts;

**SIMULATION 2:**  $\Gamma = 1$ ,  $N_H = 10^{21}\text{cm}^{-2}$ , and  $10^5$  counts;

**SIMULATION 3:**  $\Gamma = 2$ ,  $N_H = 10^{23}\text{cm}^{-2}$ , and  $10^4$  counts; and

**SIMULATION 4:**  $\Gamma = 1$ ,  $N_H = 10^{21}\text{cm}^{-2}$ , and  $10^4$  counts

respectively.

To illustrate the effect of calibration uncertainty, we selected the 15 curves in  $A_l \in \mathcal{A}$  with the largest maximum values and the 15 curves with the smallest maximum values. In some sense, these are the 30 most extreme effective area curves in  $\mathcal{A}$ . They are plotted as  $A_l - A_0$  in the first panel of Figure 2, along with a horizontal line at zero that represents the default ( $A_0 - A_0$ ). We used the Bayesian method of van Dyk et al. (2001) to fit SIMULATION 1 and SIMULATION 2 each 31 times, using each of the 31 curves of  $A_l$  plotted in Figure 2. The resulting marginal and joint posterior distributions for  $\Gamma$  and  $N_H$  appear in rows 2-4 of Figure 2; the contours plotted in the third row correspond to a posterior probability of 95% for each fit.<sup>2</sup> The figure clearly shows that the effect of calibration uncertainty swamps the ordinary statistical error. The scientist who assumes that the true effective area is known to be  $A_0$  may dramatically underestimate the error bars, and may miss the correct region entirely.

As a second illustration we fit SIMULATION 1 and SIMULATION 3 each 31 times, using the same  $A_l$  as in Figure 2 and with  $A_0$ , again using the method of van Dyk et al. (2001). The resulting posterior distributions of  $\Gamma$  and  $N_H$  are plotted in Figure 3. Comparing the two columns of the figure, the relative contribution of calibration uncertainty to the total error bars appears to grow with counts. For this reason, accounting for calibration uncertainty is especially important with rich high-count spectra. In fact, in our simulations there appears to be a limiting value where the statistical errors are negligible and the total

---

<sup>2</sup>The contours in Figure 2 were constructed by peeling (Green 1980) the original Monte Carlo sample. This involves removing the most extreme sampled values which are defined as the vertices of the smallest convex set containing the sample (i.e., the convex hull). This is repeated until only 95% of the sample remains. The final hull is plotted as the contour. This is a reasonable approximation because the posterior distributions appear roughly convex.

error bars are due entirely to calibration uncertainty. The total error bars do not go below this limiting value regardless of how many counts are observed.

We must emphasize, however, that we are assuming that the observed counts are uninformative as to which of the calibration products in the calibration sample are more or less likely. If we were not to make this assumption, however, and if a data set were so large that we were able to exclude a large portion of the calibration sample as inconsistent with the data, the remaining calibration uncertainty would be reduced and its effect would be mitigated. In this case, the default effective area and effective area curves similar to the default could potentially be found inconsistent with the data and thus the fitted model parameters could be different from what we would get if we simply relied on the default curve. In this article, however, we assume that either the data set is not large enough to be informative for the calibration products or that we do not wish to base instrumental calibration on the idiosyncrasies of a particular data set.

Both Figures 2 and 3 suggest that while the fitted values depend on the choice of  $A \in \mathcal{A}$ , the statistical errors for the parameters given any fixed  $A \in \mathcal{A}$  are more-or-less constant. The systematic errors due to calibration uncertainty shift the fitted value but do not effect its variance. Of course, in practice we do not know  $A$  and must marginalize over it, so the total error bars are larger than any of the errors bars that are computed given a particular fixed  $A$ . How to coherently compute error bars that account for calibration uncertainty is our next topic.

### 3. Spectral Analysis Using a Calibration Sample of the Effective Area

In this section, we outline how the calibration sample can be used in principled statistical analyses and describe how the complex calibration sample can be summarized in

a concise and complete manner using PCA.

### 3.1. Statistical Analysis with a Calibration Sample

#### 3.1.1. Marginalizing over Calibration Uncertainty

In a standard astronomical data analysis problem, as represented by Equation 1, it is assumed that  $A \equiv A_0$  and that  $\theta$  is estimated using  $p(\theta|Y, A_0)$ , where  $Y$  is the observed counts and  $p$  is an objective function used for probabilistic estimation and calculation of error bars. Typical choices of  $p$  are the Bayesian posterior distribution, the likelihood function, the Cash statistic, or a  $\chi^2$  statistic. We use the notation  $p(\theta|Y, A_0)$  because we generally take a Bayesian perspective, with  $p(\cdot)$  representing a probability distribution and the notation “|” referring to conditioning, e.g.,  $p(U|V)$  is to be read as “the probability of  $U$  given that  $V$  is true.”

When  $A$  is unknown, it becomes a nuisance parameter<sup>3</sup> in the model, and the appropriate objective function becomes  $p(\text{model parameters}, A|\text{data})$ . Using Bayesian notation,

$$p(\theta, A|Y, Z) = p(\theta|Y, Z, A)p(A|Y, Z),$$

where the primary source of information for  $A$  is not the observation counts,  $Y$ , but the large datasets and physical calculations used by calibration scientists, and which we denote here by  $Z$ . Generally speaking, we expect the information for  $\theta$  to come from  $Y$  rather than  $Z$ , at least given  $A$  and we expect the information for  $A$  to come from  $Z$  rather than  $Y$ .

---

<sup>3</sup>A nuisance parameter is simply an unknown but necessary parameter in the model that is not of direct interest. Its presence in the model may complicate inference, which can be a nuisance.



This can be expressed mathematically by two conditional independence assumptions:

1.  $p(\theta|Y, Z, A) = p(\theta|Y, A)$ , and
2.  $p(A|Y, Z) = p(A|Z)$ .

We make these conditional independence assumptions, and implicitly condition on  $Z$  throughout this article. In this case, we can rewrite the above equation as

$$p(\theta, A|Y) = p(\theta|Y, A)p(A), \quad (2)$$

which effectively replaces the posterior distribution  $p(A|Y)$  with the prior distribution  $p(A)$ .

Finally, we can focus attention on  $\theta$  by marginalizing out  $A$ ,

$$\begin{aligned} p(\theta|Y) &\approx \int p(\theta|Y, A)p(A)dA \\ &\approx \frac{1}{L} \sum_{l=1}^L p(\theta|Y, A_l). \end{aligned} \quad (3)$$

That is, the objective function is simply the average of the objective functions used in the standard analysis, but with  $A_0$  replaced by each of the  $A_l \in \mathcal{A}$ . Thus, the marginalization in Equation 2 does not necessarily involve estimating  $p(A|Y)$  nor specifying a parametric prior or posterior distributions for  $A$ . When this marginalization is properly computed, systematic errors from calibration uncertainty are rigorously combined with statistical errors without need for Gaussian quadrature.

Of course, when  $L$  is large as in the calibration sample of Drake et al. (2006), evaluating and optimizing Equation 3 would be a computationally expensive task. In this section we outline two strategies that aim to significantly simplify the necessary computation. The first is a general purpose but approximate strategy that can be used with any standard model fitting technique and the second is a simple adaptation that can be employed when Monte Carlo is used in Bayesian model fitting. Details and illustrations of both methods appear in §4.

### 3.1.2. Multiple Imputation

The first strategy takes advantage of a well-established statistical technique known as *multiple imputation* that is designed to handle missing data (Rubin 1987, Schafer 1997). Multiple Imputation relies on the availability of a number of Monte Carlo replications of the missing data. The replications are called the *imputations* and are designed to represent the statistical uncertainty regarding the unobserved values of the missing data. Although the calibration products are not missing data *per se*, the calibration sample provides exactly what is needed for us to apply the method of multiple imputation: a Monte Carlo sample that represents the uncertainty in an unobserved quantity.

With the calibration sample in hand, it is straightforward to apply multiple imputation. A subset of  $\mathcal{A}$  of size  $M \ll L$  is randomly selected and called the multiple imputations or the *multiple imputation sample*. The standard data analysis method is then applied  $M$  times, once with each of the  $M$  imputations of the calibration products. This produces  $M$  sets of parameter estimates along with their estimated variance-covariance matrices<sup>4</sup>, which we denote  $\hat{\theta}_m$  and  $\text{Var}(\hat{\theta}_m)$ , respectively, for  $m = 1, \dots, M$ . In the simplest form of the method of multiple imputation, we assume that each  $\hat{\theta}_m$  follows a multivariate normal distribution with mean  $\theta$ . The final fitted values and error bars are computed using a set of simple moment calculations known as the *multiple imputation combining rules* (e.g., Harel and Zhou 2005). The parameter estimate is computed simply as the average of the individual fitted values,

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m. \quad (4)$$

---

<sup>4</sup>The variance-covariance matrix is a matrix that has the square of the error bars along its diagonal and the covariance terms as off-diagonal elements. Recall that the covariance is the correlation times the product of the error bars:  $\text{Cov}(X, Y) = \text{Correlation}(X, Y)\sigma_x\sigma_Y$ .

To compute the error bars, we must combine two sources of uncertainty: the statistical uncertainty that would arise even if the calibration product were known with certainty and the systematic uncertainty stemming from uncertainty in the calibration product. Each of the  $M$  standard analyses is computed as if the calibration product were known and therefore each  $\text{Var}(\hat{\theta}_m)$  is an estimate of the statistical uncertainty. Our estimate of the statistical uncertainty is simply the average of these individual estimates,

$$W = \frac{1}{M} \sum_{m=1}^M \text{Var}(\hat{\theta}_m). \quad (5)$$

The systematic uncertainty, on the other hand, is estimated by looking at how changing the calibration product in each of the  $M$  analyses affects the fitted parameter. Thus, the systematic uncertainty is estimated as the variance of the fitted values,

$$B = \frac{1}{M-1} \sum_{m=1}^M (\hat{\theta}_m - \hat{\theta})(\hat{\theta}_m - \hat{\theta})^\top. \quad (6)$$

Finally the two components of variance are combined for the total uncertainty,

$$T = W + \left(1 + \frac{1}{M}\right) B, \quad (7)$$

where the  $\frac{1}{M}$  term accounts for small number  $M$  of imputations. If  $M$  is small relative to the dimension of  $\theta$ ,  $T$  will be unstable, and more sophisticated estimates should be used (e.g., Li et al. 1991). Here we focus on univariate summaries and error bars which depend only on one element of  $\hat{\theta}$  and the corresponding diagonal element of  $T$ .

When computing the error bars for one of the univariate fitted parameters in  $\hat{\theta}$ , say component  $m$  of  $\hat{\theta}$ , it is generally recommended that the number of sigma used be inflated to adjust for the typically small value of  $M$ . That is, rather than using one- and two-sigma for 68.3% and 95.4% intervals as is appropriate for the normal distribution, a  $t$  distribution should be used, requiring a larger number of sigma to obtain 68.3% and 95.4% intervals. In the univariate case, the *degrees of freedom* of the  $t$  distribution determine the degree of

inflation and can be estimated by

$$\text{degrees of freedom} = (M - 1) \left( 1 + \frac{MW_{mm}}{(M + 1)B_{mm}} \right)^2, \quad (8)$$

where  $W_{mm}$  and  $B_{mm}$  are the  $m$ th diagonal terms of  $W$  and  $B$ .

The method of multiple imputation is based on a number of assumptions. First, it is designed to give approximate error bars on  $\theta$  that include the effects of the imputed quantity, but if a full posterior distribution on  $\theta$  is desired, then a more detailed Bayesian calculation must be performed (see below). It will provide an approximately valid answer in general when the imputation model is compatible with the estimation procedure, i.e., when  $\hat{\theta}$  is the posterior mode from essentially the same distribution as is used for the imputation (Meng 1994). Furthermore, the computed standard deviations  $\sqrt{T}$  can be identified with 68% credible intervals only when the posterior distributions are multi-variate Normal. Additionally, when  $M$  is small, the coverage must be adjusted using the  $t$ -distribution (Equation 8).

### 3.1.3. Monte Carlo in a Bayesian Statistical Analysis

Multiple imputation offers a simple general strategy for accounting for calibration uncertainty using standard analysis methods. Because this method is only approximate, however, our preferred solution is a Monte Carlo method that is robust, reliable, and fast. In principle, Monte Carlo methods can handle any level of complexity present in both the astrophysical models and in the calibration uncertainty. Monte Carlo can be used to construct powerful methods that are able to explore interesting regions in high-dimensional parameter spaces and, for instance, determine best-fit values of model parameters along with their error bars. In this context, it is used as a fitting engine, similar to Levenberg-Marquardt, Powell, Simplex, and other minimization algorithms. One of

its main advantages is that it is highly flexible and can be applied to a wide variety of problems. A single run is sufficient to describe the variations in the model parameters that arise due to both statistical and systematic errors, which therefore leads to reduced computational costs.<sup>5</sup> Consider a Monte Carlo sample obtained by sampling the model parameters  $\theta$  given the data,  $Y$ , and the calibration product,  $A = A_0$ ,

$$\theta^{(k)} \sim p(\theta|Y, A_0),$$

where  $k$  is the iteration number and  $\theta^{(k)}$  are the values of the parameters at iteration  $k$ . The set of parameter values thus obtained is used to estimate the best-fit values and the error bars. When calibration uncertainty is included, we can no longer condition on  $A_0$  as a known value of the calibration product. Instead we add a new step that updates  $A$  according to the calibration uncertainties. In particular,  $\theta^{(k)}$  is updated using the same iterative algorithm as above, with an additional step at each iteration that updates  $A$ . Suppose at iteration  $k$ ,  $A^{(k)}$  is the realization of the calibration product. Then the new algorithm consists of the following two steps:

$$\begin{aligned} A^{(k)} & \text{ is sampled from } p(A|Y) \text{ and} \\ \theta^{(k)} & \text{ is sampled from } p(\theta|Y, A^{(k)}). \end{aligned}$$

Under the conditional independence assumptions of Section 3.1.1, we can simplify this sampler by replacing  $p(A|Y)$  with  $p(A)$  in the first step:

$$A^{(k)} \text{ is sampled from } p(A) \text{ and} \tag{9}$$

---

<sup>5</sup>In most cases, Markov chain Monte Carlo (MCMC) rather than simple Monte Carlo is required to explore complicated parameter spaces. Unfortunately, the use of MCMC in this situation raises certain technical complications. In this section we avoid these complications by focusing on the simple case of direct Monte Carlo sampling. More realistic MCMC samplers and associated complications are discussed in §4.2.

$$\theta^{(k)} \quad \text{is sampled from} \quad p(\theta|Y, A^{(k)}). \quad (10)$$

This independence assumption gives us the freedom not to estimate the posterior distribution  $p(A|Y)$  and simplifies the structure of the algorithm. It effectively separates the complex problem of model fitting in the presence of calibration uncertainties into two simpler problems: (i) fitting a model with known calibration and (ii) the quantification of calibration uncertainties independent of the current data  $Y$ .

### 3.2. Simple Summaries of a Complex Calibration Sample

The methods that we propose so far require storage of a large number of replicates of  $A \in \mathcal{A}$ . Since calibration products can be observation specific this requires a massive increase in the size of calibration databases. This concern is magnified when we consider uncertainties in the energy redistribution matrix,  $R$ , and point spread function,  $P$ , and combining multiple observations, each with their own calibration products. Although in principle this could be addressed by developing software that generates the calibration sample on the fly, we propose a more realistic and immediate solution that involves statistical compression of  $\mathcal{A}$ . Compression of this sort takes advantage of the fact that many of the replicates in  $\mathcal{A}$  differ very little from each other and in principle we can reduce the sample’s dimensionality from thousands to only a few with little loss of information. Here we describe how principal component analysis (PCA) can accomplish this for the *Chandra*/ACIS-S calibration sample generated by Drake et al. (2006) and illustrated in Figure 1.

PCA is a commonly applied linear technique for dimensionality reduction and data compression (Jolliffe 2002, Anderson 2003, Ramsay and Silverman 2005, Bishop 2007). Mathematically, PCA is defined as an orthogonal linear transformation of a set of variables such that the first transformed variable defines the linear function of the data with the

greatest variance, the second transformed variable define the linear function *orthogonal to the first* with the greatest variance, and so on. PCA aims to describe variability and is generally computed on data with mean zero. In practice, the mean of the data is subtracted off before the PCA and added back after the analysis. Computation of the orthogonal linear transformation is accomplished with the singular value decomposition of a data matrix with each variable having mean zero. This generates a set of eigenvectors that correspond to the orthogonal transformed variables, along with their eigenvalues that indicate the proportion of the variance correlated with each eigenvector. The eigenvectors with the largest values are known as the *principal components*. By selecting a small number of the largest principal components, PCA allows us to effectively summarize the variability of a large data set with a handful of orthogonal eigenvectors and their corresponding eigenvalues.

Our aim is to effectively compress  $\mathcal{A}$  using PCA. Using the singular vector decomposition of a matrix with rows equal to the  $A_l - \bar{A}$  with  $\bar{A} = \frac{1}{L} \sum_l A_l$ , we compute the eigenvectors  $(v_1, \dots v_L)$  and corresponding eigenvalues  $(r_1^2, \dots r_L^2)$ , ordered such that  $r_1 \geq r_2 \geq \dots \geq r_L$ . The fraction of the variance of  $\mathcal{A}$  in the direction of  $v_l$  is

$$f_l = \frac{r_l^2}{\sum_{j=1}^L r_j^2}. \quad (11)$$

In practice, this gives us the option of using a smaller number of components,  $J < L$  in the reconstruction, that is sufficient to account for a certain fraction of the total variance. A large amount of compression can be achieved because very few components are needed to compute the effective area to high precision. For example, in the case of ACIS effective areas, 8-10 components (out of 1000) can account for 95% of the variance, and  $\approx 20$  components can account for 99% of the variance. Note that this approximation is valid only when considered over the full energy range; small localized variations in  $\mathcal{A}$  that contribute little to the total variance, even if they may play a significant role in specific analyses (the depth of the C-edge, for example) may not be accounted for.

With the PCA representation of  $\mathcal{A}$  in hand, we wish to generate replicates of  $A$  that mimic  $\mathcal{A}$ . In doing so, however, we must account for the fact that calibration products typically vary from observation to observation to reflect deterioration of the telescope over time and other factors that vary among the observations. However, even though the magnitudes of the calibration products may change, the underlying uncertainties are less variant and are comparable across different regions of the detector at different times. We thus suppose that the differences among the calibration samples can be represented by simply changing the default calibration product, at least in many cases. That is, we assume that the distribution in the calibration samples differ only in their (loosely defined) average and that differences in their variances can be ignored. Under this assumption, we can easily generate calibration replicates based on the first  $J$  principal components as

$$A^{\text{rep}} = \bar{A} + (A_0^* - A_0) + \sum_{j=1}^J e_j r_j v_j + \xi e_{J+1}, \quad (12)$$

$$= A_0^* + \delta \bar{A} + \sum_{j=1}^J e_j r_j v_j + \xi e_{J+1} \quad (13)$$

where  $A_0^*$  is the observation-specific effective area that would currently be created by users,  $A_0$  is the nominal default effective area from calibration,  $\delta \bar{A} = \bar{A} - A_0$ ,  $\xi = \sum_{j=J+1}^L r_j v_j$ , and  $(e_1, \dots, e_{J+1})$  are independent standard normal random variables. In addition to the first  $J$  principal components, this representation aims to improve the replicates by including the residual sum of the remaining  $L - J$  components. Equation 12 shows how we account for  $A_0^*$ . If  $A_0^*$  were equal to  $A_0$ , Equation 12 would reduce to the standard PCA representation. To account for the observation-specific effective area, we add the offset  $A_0^* - A_0$ . Equation 13 rearranges the terms to express  $A^{\text{rep}}$  as the sum of calibration quantities that we propose to provide in place of  $\mathcal{A}$ . In particular, using Equation 13, we can generate any number of Monte Carlo replicates from  $\mathcal{A}$ , using only  $\delta \bar{A}$ ,  $A_0^*$ ,  $(r_1 v_1, \dots, r_L v_L)$ , and  $\xi$ . In this way we need only provide instrument-specific and not observation-specific values of  $(r_1 v_1, \dots, r_L v_L)$ , and  $\xi$ .



Figure 4 illustrates the use of PCA compression on the calibration sample generated by Drake et al. (2006) and illustrated in Figure 1. We generated 1000 replicate effective areas using Equation 13 with  $A_0 = A_0^*$  and  $J = 8$ . The dashed and dotted lines in the upper left panel respectively superimpose the full range and 68.3% intervals of these replicates on the corresponding intervals for the original calibration sample, plotted in light and dark grey. In this case, using  $J = 8$  captures 96% of the variation in  $\mathcal{A}$ , as computed with Equation 11. The remaining three panels give cross sections at 1.0, 1.5, and 2.5 keV. The distributions of the 1000 replicates generated using Equation 13 appears as solid lines, and those of the original calibration sample as a gray regions. The figure shows that PCA replicates generated with  $J = 8$  are quite similar to the original calibration sample.

Although the PCA representation cannot be perfect (e.g., it does not fully represent uncertainty overall or in certain energy regions) it is much better than not accounting for uncertainty at all.

#### 4. Algorithms Accounting for Calibration Uncertainty

In this section we describe specific algorithms that incorporate calibration uncertainty into standard data analysis routines. In §4.1 we show how multiple imputation can be used with popular scripted languages like *HEASARC/XSPEC* and *CIAO/Sherpa* for spectral fitting, and in §4.2 we describe some minor changes that can be made to sophisticated Markov chain Monte Carlo samplers to include the calibration sample. In both sections we begin with cumbersome but precise algorithms and then show how approximations can be made to simplify the implementation. Our recommended algorithms appear in §4.1.2 and §4.2.2. In §5 we demonstrate that these approximations have a negligible effect on the final fitted values and error bars.

## 4.1. Algorithms for Multiple Imputation

### 4.1.1. Using the Full Calibration Sample

Multiple imputation is an easy to implement method that relies heavily on standard fitting routines. An algorithm for accounting for calibration uncertainty using multiple imputation can be described by:

STEP 1: For  $m = 1, \dots, M$ , repeat the following:

STEP 1A: Randomly sample  $A_m$  from  $\mathcal{A}$ .

STEP 1B: Fit the spectral model (e.g., using *Sherpa*) in the usual way, but with effective area set to  $A_m$

STEP 1C: Record the fitted values of the parameters as  $\hat{\theta}_m$

STEP 1D: Compute the variance-covariance matrix of the fitted values and record the matrix as  $\text{Var}(\hat{\theta}_m)$ . (In *Sherpa* this can be done using the `covariance` function.)

STEP 2: Use Equation 4 to compute the fitted value,  $\hat{\theta}$  of  $\theta$ .

STEP 3: Use Equations 5–7 to compute the variance-covariance matrix,  $\text{Var}(\hat{\theta}) = T$ , of  $\hat{\theta}$ .

The square root of the diagonal terms of  $\text{Var}(\hat{\theta}) = T$  are the error bars of individual parameters.

STEP 4: Use Equation 8 to compute the degrees of freedom for each component of  $\hat{\theta}$  which are used to properly calibrate the error bars computed in STEP 3.

Asymptotically,  $\pm 1\sigma$  error bars correspond to equal-tail 68.3% intervals under the Gaussian distribution. When the number of imputations is small,  $\pm t_{\text{df}}\sigma$  error bars should be used instead, where  $t_{\text{df}}$ , a number  $> 1$ , can be looked up in any standard  $t$ -distribution table using “df” equal to the degrees of freedom computed in STEP 4, see §5.1 for an illustration.

If the correlations among the fitted parameters are not needed, the error bars of the individual fitted parameters can be computed one at a time using Equations 5–7 with  $\hat{\theta}_m$  and  $\text{Var}(\hat{\theta}_m)$  replaced by the fitted value of the individual parameter and the square of its error bars, both computed using  $A_m$ .

#### 4.1.2. Using the PCA Approximation

Using the PCA approximation results in a simple change to the algorithm in §4.1.2: STEP 1A is replaced by (see Equation 13):

STEP 1A: Set  $A_m = A_0^* + \delta\bar{A} + \sum_{j=1}^J e_j r_j v_j + \xi e_{J+1}$ , where  $(e_i, \dots, e_{J+1})$  are independent standard normal random variables.

The choice between this algorithm and the one described in Section 4.1.1 should be determined by the availability of a sample of size  $M$  from  $A$  (in which case the Algorithm in Section 4.1.1 should be used) or of the PCA summaries of  $A$  required for the algorithm in this section.

## 4.2. Algorithms for Monte Carlo in a Bayesian Analysis

In §3.1.3 we considered simple Monte Carlo methods that simulate directly from the posterior distribution,  $\theta^{(k)} \sim p(\theta|Y, A_0)$ . More generally, Markov chain Monte Carlo (MCMC) methods can be used to fit much more complicated models. (Good introductory references to MCMC can be found in Gelman 2003 and Gregory 2005.) A Markov chain is an ordered sequence of parameter values such that any particular value in the sequence depends on the history of the sequence only through its immediate predecessor. In this way MCMC samplers produce dependent draws from  $p(\theta|Y, A_0)$  by simulating

$\theta^{(k)}$  from a distribution that depends on the previous value of  $\theta$  in the Markov chain,  $\theta^{(k)} \sim \mathcal{K}(\theta|\theta^{(k-1)}; Y, A_0)$ . That is,  $\mathcal{K}$  is designed to be simple to sample from, while the full  $p(\theta|Y, A_0)$  may be quite complex. The price of this, however, is that the  $\theta^{(k)}$  may not be statistically independent of the  $\theta^{(k-1)}$ ; and in fact may have appreciable correlation with  $\theta^{(k-d)}$  (that is, an autocorrelation of length  $d$ ). The distribution  $\mathcal{K}$  is derived using methods such as the Metropolis-Hastings algorithm and/or the Gibbs sampler that ensures that the resulting Markov chain converges properly to  $p(\theta|Y, A_0)$ . Van Dyk et al. (2001) show how Gibbs sampling can be used to derive  $\mathcal{K}$  in high-energy spectral analysis. Their method has recently been generalized in a *Sherpa* module called **pyBLoCXS** (Bayesian Low Count X-ray Spectral analysis in Python, to be released)<sup>6</sup> In this section we show how **pyBLoCXS** can be modified to account for calibration uncertainty. For clarity we use the notation

$$\theta^{(k)} \sim \mathcal{K}_{\text{pyB}}(\theta|\theta^{(k-1)}; Y, A) \quad (14)$$

to indicate a single iteration of **pyBLoCXS** run with the effective area set to  $A$ .

#### 4.2.1. A Pragmatic Bayesian Method

In §3.1.3 we describe how a Monte Carlo sampler can be constructed to account for calibration uncertainty under the assumption that the observed counts carry little information as to the choice of effective area curve. In particular, we must iteratively update  $A^{(k)}$  and  $\theta^{(k)}$  by sampling them as described in Equations 9 and 10. Sampling  $A^{(k)}$  from  $p(A)$  can be accomplished by simply selecting an effective area curve at random from  $\mathcal{A}$ . Updating  $\theta$  is more complicated, however, because we are using MCMC. We cannot

---

<sup>6</sup>URL: <http://cxc.harvard.edu/sherpa>. The **pyBLoCXS** routine uses a different choice of  $\mathcal{K}$  that relies more heavily on Metropolis-Hastings than on Gibbs sampling and can accommodate a larger class of spectral models.

directly sample  $\theta^{(k)}$  from  $p(\theta|Y, A^{(k)})$  as stipulated by Equation 10. The `pyBLoCXS` update of  $\theta^{(k)}$  depends on the previous iterate,  $\theta^{(k-1)}$ . Thus, we must iterate STEP 2 of the fully Bayesian sampler several times before it converges and delivers an uncorrelated draw from  $p(\theta|Y, A^{(k)})$ . In this way, we iterate STEP 2 in the following sampler until the dependence on  $\theta^{(k-1)}$  is negligible. To simplify notation, we display iteration  $k + 1$  rather than iteration  $k$ ; notice that after  $I$  repetitions, STEP 2 returns  $\theta^{(k+1)}$ . In practice we expect a relatively small value of  $I$  ( $\sim 10$  or fewer) will be sufficient, see §5.2. The MCMC step for a given  $k$  is as follows:

STEP 1: Sample  $A^{(k+1)} \sim p(A)$ .

STEP 2: For  $i = 1, \dots, I$ ,

$$\text{Sample } \theta^{(k+i/I)} \sim \mathcal{K}_{\text{pyB}}(\theta|\theta^{(k+(i-1)/I)}; Y, A^{(k+1)}).$$

Once the MCMC sampler run is completed, the ‘best-fit’ and confidence bounds for each parameter are typically determined from the mean and widths of the histograms constructed from the traces of  $\{\theta^k\}$ ; or mean and widths of the contours (for multiple parameters), as in Figures 2 and 7; see Park et al. (2008) for discussion.

#### 4.2.2. A Pragmatic Bayesian Method with the PCA Approximation

Using the PCA approximation results in a simple change to the algorithm in §4.2.1:

STEP 1 is replaced by

STEP 1: Set  $A^{(k+1)} = \delta\bar{A} + A_0^* + \sum_{j=1}^J e_j r_j v_j + \xi e_{J+1}$ , where  $(e_1, \dots, e_{J+1})$  are independent standard normal random variables.

Because of the advantages in storage that this method confers, and the negligible effect that the approximation has on the result (see §5.3), this is our recommended method when using MCMC to account for calibration uncertainty with data sets with ordinary counts.

## 5. Numerical Evaluation

In this section we investigate optimal values of the tuning parameters needed by the algorithms and compare the performance of the algorithms with simulated and with real data. Throughout, we use the absorbed power law simulations described in Table 2 to illustrate our methods and algorithms. The eight simulations represent a  $2 \times 2 \times 2$  design with the three factors being (1) data simulated with  $A_0$  and with an extreme effective area curve from  $\mathcal{A}$ , (2)  $10^5$  and  $10^4$  nominal counts, and (3) two power law spectral models. These simulations include the four described in §2.2. We investigate the number of imputations required in Multiple Imputation studies in §5.1, and the number of subiterations required in MCMC runs in §5.2. We compare the results from the different algorithms (Multiple Imputation with sampling and with PCA, and `pyBLcXS` with sampling and PCA) in detail in §5.3, and apply them to a set of Quasar spectra in §5.4.

### 5.1. Determining the Number of Imputations

When using multiple imputation, we must decide how many imputations are required to adequately represent the variability in  $\mathcal{A}$ . Although in the social sciences as few as 3-10 imputations are sometimes recommended (e.g., Schafer 1997), larger numbers more accurately represent uncertainty. To investigate this we fit spectra from SIMULATION 1 and SIMULATION 2 using *Sherpa*, with different values of  $M$ , the number of imputations. For each value of  $M$  we generate  $M$  effective area curves,  $A^{\text{rep}}$ , using Equation 13, fit the

simulated spectrum  $M$  times, once with each  $A^{\text{rep}}$ , derive the  $1\sigma$  error bars, and combine the  $M$  fits using the multiple imputation combining rules in Equations 4–7. This gives us a single total error bar for each parameter. We repeat this process 200 times for each value of  $M$  to investigate the variability of the computed error bar for each value of  $M$ . The result appears in the first two rows of Figure 5. For small values of  $M$  the error bars are often too small or too large. With  $M$  larger than about 20, however, the multiple imputation error bars are quite accurate. Even with  $M = 2$ , however, the error bars computed with multiple imputation are more representative of the actual uncertainty than when we fix the effective area at  $A_0$ , which is represented by  $M = 1$  in Figure 5. Generally speaking,  $M = 20$  is usually adequate, but  $M = 20$  to 50 is better if computational time is not an issue. Note that the size of the calibration sample  $\mathcal{A}$  is generally much larger than this, and it is therefore a fair sample to use in the Bayesian sampling techniques described in §4.2.

When  $M$  is relatively small, the computed  $\pm 1\sigma$  error bars may severely underestimate the uncertainty, and must be corrected for the degrees of freedom in the imputations (see Equation 8). To illustrate this, we compute the nominal coverage of the standard  $\pm \text{one}\sqrt{T}$  interval for each of the MI analyses described in the previous paragraph. When  $M$  is large, such intervals are expected to contain the true parameter value 68.3% of the time, the probability that a Gaussian random variable is within one standard deviation of its mean. With small  $M$ , however, the coverage decreases because of the extra uncertainty in the error bars. The bottom two rows of Figure 5 illustrate the importance of adjusting for the degrees of freedom, especially when using relatively small values of  $M$ . The plots give the range of nominal coverage rates for one  $\sqrt{T}$  error bars. For large  $M$  the coverage approaches 95%, but for small  $M$  it can be as low as 50-60%. This can be corrected by computing the degrees of freedom using Equation 8 and using  $\pm t_{\text{df}}\sigma$  instead of  $\pm \text{one}\sqrt{T}$ , as described in §4.1.1.

## 5.2. Determining the Number of Subiterations in the Pragmatic Bayesian Method

As noted in §4.2.1, in order to obtain a sample from the  $\theta^{(t)} \sim p(\theta|Y, A^{(t)})$  as in Equation 10 we must iterate `pyBLoCXS`  $I$  times to eliminate the dependence of  $\theta^{(k-1)}$ . To investigate how large  $I$  must be, we run `pyBLoCXS` on the spectra from SIMULATIONS 1 and SIMULATION 5 of Table 2, which were generated using the “default” and an “extreme” effective area curve. Since SIMULATION 5 was generated using the “extreme” effective area curve, it is the “extreme” curve that is actually “correct” and the “default” curve that is “extreme”. When running `pyBLoCXS` with the “default” effective area curve, we initiated the chain at the posterior mean of the parameters given the “extreme” curve, and vis versa. This ensures that we are using a relatively extreme starting value and will not underestimate how large  $I$  must be to generate an essentially independent draw. The resulting autocorrelation and time series plots for  $\Gamma$  appear in Figure 6. The autocorrelation plots report the correlation of  $\theta^{(k)}$  and  $\theta^{(k+I)}$  for each value of  $I$  plotted on the horizontal axis. The plots show that for  $I > 10$  the autocorrelations are essentially zero for both spectra, and we can consider  $\theta^{(k)}$  and  $\theta^{(k+10)}$  to be essentially independent. Similarly, the time series plots show that there is no effect of the starting value past the tenth iteration. Similar plots for  $N_{\text{H}}$  and the normalization parameter (not included) are essentially identical. Thus, in all subsequent computations we set  $I = 10$  in the pragmatic Bayesian samplers. Generally speaking, the user should construct autocorrelation plots to determine how large  $I$  must be in a particular setting.

When we iterate Step 2 in the Pragmatic Bayesian Method, we are more concerned with the mixing of the chain once it has reached its stationary distribution, rather than convergence of the chain to its stationary distribution. This is because convergence to the stationary distribution will be assessed using the final chain of  $\theta^{(t)}$  in the regular way,



i.e., using multiple chains (Gelman & Rubin 1992, van Dyk et al. 2001). Even after the stationary distribution has been reached, we need to obtain a value of  $\theta^{(t+1)}$  in Step 2 that is essentially independent of the previous draw, given  $A^{(k+1)}$ . Thus, we focus on the autocorrelation of the chain  $\theta^{(t)}$  for fixed  $A$ . This said, if the posterior of  $\theta$  is highly dependent on  $A$  and  $A^{(t)}$  and  $A^{(t+1)}$  are extreme within the calibration sample, that the conditional posterior distribution of  $\theta$  given  $A^{(t)}$  and  $A^{(t+1)}$  may be quite different and we may need to allow  $\theta$  to converge to its new conditional posterior distribution. The time series plots in Figure 6 investigate this possibility when extreme values of  $A$  are used. Luckily, the effect of these extreme starting values still burns off in just a few iterations, as is evident in Figure 6.

### 5.3. Comparing the Algorithms

We discuss two classes of algorithms in §4 to account for calibration uncertainty in spectral analysis: Multiple Imputation, and a pragmatic Bayesian MCMC sampler. For each, we consider two methods of exploring the calibration product sample space: first by directly sampling from the set of effective areas  $\mathcal{A}$ , and second by simulating an effective area from a compressed Principal Component representation. Here, we evaluate the effectiveness of each of the four resulting algorithms, and show that they all produce comparable results, and are a significant improvement over not including the calibration uncertainty in the analysis. We fit each of the eight simulated data sets described in Table 2 using each of the four algorithms. The first four simulations are identical to those described in §2.2. Analyses carried out using Multiple Imputation all used  $M = 20$  imputations. For analyses using the PCA approximation to  $\mathcal{A}$ , we used  $J = 17$ . For pragmatic Bayesian methods, we used  $I = 10$  inner iterations. Figure 7 gives the resulting estimated marginal posterior distributions for  $\Gamma$  for each of the eight simulations and each of the four fitting

algorithms along with the results when the effective area is fixed at  $A_0$ . Parameter traces (also known as time series) are also shown for all the simulations for the two MCMC algorithms (see §4.2). Although the fitted values differ somewhat (see Simulations 1, 2, 3, and 6) among the four algorithms that account for calibration uncertainty, the differences are very small relative to the errors and overall the four methods are in strong agreement. When we do not account for calibration uncertainty, however, the error bars can be much smaller and in some cases the nominal 68% intervals do not cover the true value of the parameter (see Simulations 1, 2, 5, and 6, corresponding to larger nominal counts). When we do account for calibration uncertainty, only in Simulation 6 did the 68% intervals not contain the true value, and in this case the 95% (not depicted) do contain the true value. Results for  $N_H$  are similar but omitted from Figure 7 to save space.

An advantage of using MCMC is that it maps out the posterior distribution (under the conditional independence assumptions of Section 3.1.1) rather than making a Gaussian approximation to the posterior distribution. Notice the non-Gaussian features in the posterior distributions plotted for Simulations 1, 3, 5, and 7 (corresponding to the harder spectral model).

#### 5.4. Application to a Sample of Radio Loud Quasars

Here we illustrate our methods with a realistic case, using X-ray spectra available for a small sample of radio loud quasars observed with the *Chandra* X-ray Observatory in 2002 (Siemiginowska et al. 2008). We performed the standard data analysis including source extraction and calibration with CIAO software (*Chandra* Interactive Analysis of Observations). The X-ray emission in radio loud quasars originates in a close vicinity of a supermassive black hole and could be due to an accretion disk or a relativistic jet. It is well described by a Compton scattering process and the X-ray spectrum can be modeled by an

absorbed power law:

$$S(E) = N E^{-\Gamma} e^{-\sigma(E) N_{\text{H}}} \text{ photons cm}^{-2} \text{ sec}^{-1} \text{ keV}^{-1}, \quad (15)$$

where  $\sigma(E)$  is the absorption cross-section, and the three model parameters are: the normalization at 1 keV,  $N$ ; the photon index of the power law,  $\Gamma$ ; and the absorption column,  $N_{\text{H}}$ .

The number of counts in the X-ray spectra varied between 8 and 5500. After excluding two datasets (ObsID 3099 which had 8 counts, and ObsID 836 which is better described by a thermal spectrum), we reanalyzed the remaining 15 sources to include calibration uncertainty. In fitting each source, we included a background spectrum extracted from the same observation over a large annulus surrounding the source region. We adopted a complex background model (a combination of a polynomial and 4 gaussians) that was first fit to the blank-sky data provided by the *Chandra* X-ray Center to fix its shape. While fitting the models to the source and background spectra, we only allow for the normalization of the background model to be free. This is an appropriate approach for very small background counts in the Chandra spectra of point sources. We used this background model for all spectra (except for two – ObsIDs 3101 and 3106 – that had short 5 ksec exposure times and small number of counts  $< 45$ , for which the background was ignored). The original analysis (Siemiginowska et al. 2008) did not take into account calibration errors, and as we show below the statistical errors are significantly smaller than the calibration errors for sources with a large number of counts.

We fit each spectrum accounting for uncertainty in the effective area in two ways:

1. with the multiple imputation method in §4.1.2 using Sherpa for the individual fits, and
2. with the pragmatic Bayesian algorithm in §4.2.2 using pyBLcXS for MCMC sampling.

Both of these fits use the PCA approximation using 14 observation-specific default effective area curves,  $A_0^*$  in Equation 13 with  $J = 17$ . We use  $M = 20$  multiple imputations and  $I = 10$  subiterations in the pragmatic Bayesian sampler. To illustrate the effect of accounting for calibration uncertainty, we compared the first fit with the Sherpa fit that fixes the effective area at  $A_0^*$  and each of the second and third fits with the pyBL0CXS fit that also fixes the effective area at  $A_0^*$ .

The results appear in Figure 8 which compares the error bars computed with ( $\sigma_{\text{tot}}$ ) and without ( $\sigma_{\text{stat}}$ ) accounting for calibration uncertainty. The left panel uses *Sherpa* and computes the total error using multiple imputation, and the right panel uses pyBL0CXS and computes the total error using the pragmatic Bayesian method. The plots demonstrate the importance of properly accounting for calibration uncertainty in high-counts, high-quality observations. The systematic error becomes prominent with high counts because the statistical error is small, and  $\sigma_{\text{tot}}$  deviates from  $\sigma_{\text{stat}}$ , asymptotically approaching a value of  $\sigma_{\text{tot}} \approx 0.04$ . This asymptotic value represents the limiting accuracy of any observation carried out with this instrument, regardless of source strength or exposure duration. For the absorbed power law model applied here, the systematic uncertainty on  $\Gamma$  becomes comparable to the statistical error for spectra with counts  $\gtrsim 2400$ , with the largest correction seen in ObsID 866, which had  $> 14500$  counts.

## 6. Discussion

In the previous sections, we have worked through a specific example (Chandra effective area) in some detail. Now, in this section, we present two more complete generalizations. The first is the case ignored previously, when the data have something interesting to say about the calibration uncertainties. In the second, we explain how to generalize the algorithms we worked through earlier to the full range of instrument responses, including

energy redistribution matrices and point spread functions .

### 6.1. A Fully Bayesian Method

To avoid the assumption that the observed counts carry little information as to the choice of effective area curve, we can employ a fully Bayesian approach that bases inference on the full posterior distribution  $p(\theta, A|Y)$ . To do this via MCMC, we must construct a Markov chain with stationary distribution  $p(\theta, A|Y)$ , which can be accomplished by iterating a two-step Gibbs sampler, for  $k = 1, \dots, K$ .

#### A Fully Bayesian Sampler

STEP 1: Sample  $A^{(k+1)} \sim p(A|\theta^{(k)}, Y)$ .

STEP 2: Sample  $\theta^{(k+1)} \sim \mathcal{K}_{\text{pyB}}(\theta|\theta^{(k)}; Y, A^{(k+1)})$ .

Notice that unlike in the pragmatic Bayesian approach in §4.2, STEP 1 of this sampler requires  $A$  to be updated given the current data. Unfortunately, sampling  $p(A|\theta^{(k)}, Y)$  is computationally quite challenging. The difficulty arises because the fitted value of  $\theta$  can depend strongly on  $A$ . That is, calibration uncertainty can have a large effect on the fitted model, see Drake et al. (2006) and §2.2. From a statistical point of view, this means that given  $Y$ ,  $\theta$  and  $A$  can be highly dependent and  $p(A|\theta^{(k)}, Y)$  can depend strongly on  $\theta^{(k)}$ . Thus a large proportion of the replicates in  $\mathcal{A}$  may have negligible probability under  $p(A|\theta^{(k)}, Y)$  and it can be difficult to find those that have appreciable probability without doing an exhaustive search. The computational challenges of a fully Bayesian approach are part of the motivation behind our recommendation of the pragmatic Bayesian method. Despite the computational challenges, there is good reason to pursue a Fully Bayesian Sampler. Insofar as the data are informative as to which replicates in  $\mathcal{A}$  are more – or

less – likely, the dependence between  $\theta$  and  $A$  can help us to eliminate possible values of  $\theta$  along with replicates in  $\mathcal{A}$ , thereby reducing the total error bars for  $\theta$ . Work to tackle the computational challenges of the fully Bayesian approach is ongoing.

## 6.2. General Methods for Handling Calibration Uncertainties

In general, the response of a detector to incident photons arriving at time  $t$  can be written as

$$M(E^*, \mathbf{x}^*, t; \theta) = \int dE d\mathbf{x} S(E, \mathbf{x}, t; \theta) R(E, E^*, \mathbf{x}^*; t) P(\mathbf{x}, \mathbf{x}^*, E; t) A(E, \mathbf{x}^*; \mathbf{x}, t) \quad (16)$$

where  $\mathbf{x}^*$  and  $E^*$  are the measured photon location and energy (or the detector channel), while  $\mathbf{x}$  and  $E$  are the true photon sky location and energy; the source physical model  $S(E, \mathbf{x}, t; \theta)$  describes the energy spectrum, morphology (point, extended, diffuse, etc.), and variability with parameters  $\theta$ ; and  $M(E^*, \mathbf{x}^*, t; \theta)$  are the expected counts in detector channel space. Calibration is carried out using well known instances of  $S(E, \mathbf{x}, t; \theta)$  to determine the quantities

$$\begin{aligned} R(E, E^*, \mathbf{x}^*; t) &\equiv \text{Energy Redistribution} \\ P(\mathbf{x}, \mathbf{x}^*, E; t) &\equiv \text{Point Spread Function} \\ A(E, \mathbf{x}^*; \mathbf{x}, t) &\equiv \text{Effective Area} \end{aligned} \quad (17)$$

It is important to note that all of the quantities in Equation 16 have uncertainties associated with them. Our goal is providing a fast, reliable, and robust strategy to incorporate the jittering patterns in all of the calibration products and to draw proper inference, best fits and error bars, reflecting calibration uncertainty.

In principle, using a calibration sample to represent uncertainty and the statistical methods for incorporating the calibration sample described in §3 and §4 can be applied

directly to calibration uncertainty for any of the calibration products. The use of PCA, however, to summarize the calibration sample may not be robust enough for higher dimensional and more complex calibration products. More sophisticated image analysis techniques or hierarchically applied PCA may be more appropriate. Our basic strategy, however, of providing instrument-specific summaries of the variability in the calibration uncertainty and observation-specific measures of the mean (or default) calibration product, is quite general. Thus, in this section, we focus on the generalization of Equation 12 and begin by rephrasing the equation as

$$\text{Replicate Calibration Product} = \text{Mean} + \text{Offset} + \text{Explained Variability} + \text{Residual Variability} . \quad (18)$$

Here the Mean is the mean of the calibration sample, the Offset is the shift that we impose on the center of distribution of the calibration uncertainty to account for observation-specific differences, the Explained Variability is the portion of the variability that summarize in parametric and/or systematic way (e.g., using PCA), and the Residual Variability is the portion of the variability left unexplained by the systematic summary. These four terms correspond to the four terms in Equation 12.

The formulation in Equation 18 removes the necessity of depending solely on PCA to summarize variance in the calibration sample, and allows us to use a variety of methods to generate the simulated calibration products. For example, we can even include such loosely stated measures of uncertainty as “the effective area is uncertain by X% at wavelength Y”. This formulation is not limited to describing effective areas alone, but can also be used to encompass the calibration uncertainty in response matrices and point spread functions. The precise method by which the variance terms are generated may vary widely, but in all foreseeable cases they can be described as in Equation 18, with an offset term and a random variance component added to the mean calibration product, and with an optional residual component. The calibration sample simulated in this way form an informative

prior  $p(A, R, P)$  that could be used like  $p(A)$  in Equation 9. Some potential methods of describing the variance terms are:

1. When a large calibration sample is available, the random component is simply the full set of calibration products in the sample. When using a Monte Carlo for model fitting, as in §3.1.3, a random index is chosen at each iteration and the calibration product corresponding to that index is used for that iteration. This process preserves the weights of the initial calibration sample. In this scenario the residual component is identically zero.
2. If the calibration uncertainty is characterized by a multiplicative polynomial term in the source model, the explained variance component in Equation 18 can be obtained by sampling the parameters of the polynomial, from a Gaussian distribution, using their best-fit values and the estimated errors. These simulated calibration products can then be used to modify the nominal products inside each iteration. Thus, the offset and residual terms are zero, and only the polynomial parameter best-fit values and errors need to be stored.
3. If a calibration product is newly identified, it may be systematically off by a fixed but unknown amount over a small passband, and users can specify their own estimate of calibration uncertainty as a randomized additive constant term over the relevant range. This is essentially equivalent to using a correction with a first-order polynomial. The stored quantities are the average offset, the bounds over which the offset can range, and a pointer specifying whether to generate uniform or Gaussian deviates over that range.



## 7. Summary

We have developed a method to handle in a practical way the effect of uncertainties in instrument response on astrophysical modeling, with specific application to *Chandra*/ACIS instrument effective area. Our goal has been to obtain realistic error bars on astrophysical source model parameters that include both statistical and systematic errors. For this purpose, we have developed a general and comprehensive strategy to describe and store calibration uncertainty and to incorporate them into data analysis. Starting from the full, precise, but cumbersome objective-function of the parameters, data, and instrument uncertainties, we adopt a Bayesian posterior-probability framework and simplify it in a few key places to make the problem tractable. This work holds practical promise for a generalized treatment of instrumental uncertainties in not just spectra but also imaging, or any kind of higher-dimensional analyses; and not just X-rays, but across wavelengths and even to particle detectors. Our scheme treats the possible variations in calibration as an informative prior distribution while estimating the posterior probability distributions of the source model parameters. Thus, the effects of calibration uncertainty is automatically included in the result of a single fit. This is different from a usual sensitivity study in that we provide an actual uncertainty estimate. Our analysis shows that systematic error contribution in high counts spectra is more significant than when there are few counts; therefore, including calibration uncertainty in a spectral fitting strategy is highly recommended for high quality data.

We adopt the calibration uncertainty variations, in particular the effective area variations for the *Chandra*/ACIS-S detector, described by Drake et al. (2006), as an exemplar case. Using the effective area sample  $\mathcal{A}$  simulated by them, we

1. show that variations in effective areas lead to large variations in fitted parameter values;

2. demonstrate that systematic errors are relatively more important for high counts, when statistical errors are small;
3. describe how the calibration sample can be effectively compressed and summarized by a small number of components from a Principal Components Analysis;
4. outline two separate algorithms with which to incorporate systematic uncertainties within spectral analysis:
  - (a) an approximate, but quick method based on the Multiple Imputation combining rule that carries out spectral fits for different instances of the effective area and merges the mean of the variances with the variance of the means; and
  - (b) a pragmatic Bayesian method that incorporates sampling of the effective areas as from a prior distribution within an MCMC iteration scheme.
5. detail two methods of sampling  $A^{\text{rep}}$ : directly from the calibration sample  $\mathcal{A}$ , and via a PCA decomposition
6. show that  $\approx 20$  representative samples of  $A^{\text{rep}}$  are needed to obtain relatively reliable estimates of uncertainty;
7. apply the method to a real dataset of a sample of quasars and show that known systematic uncertainties require that, e.g., the power-law index  $\Gamma$  cannot be determined with an accuracy better than  $\sigma_{\text{tot}}(\Gamma) \approx 0.04$ ; and
8. discuss future directions of our work, both in relaxing the constraint of not allowing the calibration sample  $\mathcal{A}$  to be affected by the data, and in generalizing the technique to other sources of calibration uncertainty.

This work was supported by NASA AISRP grant NNG06GF17G (HL, AC, VLK), and CXC NASA contract NAS8-39073 (VLK, AS, JJD, PR), NSF grants DMS 04-06085 and

DMS 09-07522 (DvD, AC, SM, TP), and NSF grants DMS-0405953 and DMS-0907185 (XLM). We acknowledge useful discussions with Herman Marshall, Alex Blocker, Jonathan McDowell, and Arnold Rots.

## REFERENCES

- Aguirre, et al., 2011, *ApJS*, 192, 4
- Anderson, T.W., 2003, *An Introduction to Multivariate Statistical Analysis*, 3<sup>rd</sup> ed., John Wiley & Sons, NY
- Arnaud, K. A., 1996, *Astronomical Data Analysis Software and Systems V*, 101, 17
- Bevington, P.R., and Robinson, D.K., 1992, *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, 2<sup>nd</sup> ed.
- Bishop, C., 2007, *Pattern Recognition and Machine Learning*, 1<sup>st</sup> ed., Springer, NY
- Bridle, S.L., et al., 2002, *MNRAS*, 335(4), 1193
- Brown, A., 1997, *The Neutron and the Bomb: A biography of Sir James Chadwick*, Oxford University Press
- Butler, R. P., Marcy, G. W., Williams, E., McCarthy, C., Dosanjk, P., & Vogt, S. S. 1996, *PASP*, 108, 500
- Casella, G., and Berger, R.L., 2001, *Statistical Inference*, 2<sup>nd</sup> ed., Duxbury Press, CA
- Christie, M.A., et al., 2005, *Los Alamos Science* 29, 6
- Conley, et al., 2011, *ApJS*, 192, 1, 1
- Cox, M.G., and Harris, P.M., 2006, *Meas. Sci. Technol.*, 17, 533
- David, L., et al., 2007, *Chandra Calibration Workshop*, #2007.23
- Davis, J.E., 2001, *ApJ*, 548, 1010
- Drake, J.J., et al., 2006,
- Ferraty, F., and Vieu, P., 2006, *Nonparametric Functional Data Analysis: Theory and Practice*, Springer, 1<sup>st</sup> ed., NY
- Forrest, D.J. et al., 1997, Technical Report, New Hampshire Univ. Durham, NH

- Forrest, D.J., 1988, BAAS, 20, p. 740
- Freeman, P., et al., 2001, Proc. SPIE, 4477, 76
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., 2003, *Bayesian Data Analysis*, Second Edition, Chapman & Hall/CRC Texts in Statistical Science
- Gelman, A., and Rubin, D.B., 1992, Statistical Sci., 7, 457
- Green, P.J., 1980, *Interpreting Multivariate Data*, 3-19. Chinchester: Wiley. p241
- Gregory, P.C., 2005, *Bayesian Logical Data Analysis for the Physical Sciences*, in *X-Ray Astronomy Handbook*, Cambridge University Press
- Grimm, H.-J., et al., 2009,
- Hanlon, L.O., et al., 1995, Ap&SS, 231, 157
- Harel, O., and Zhou, X. H. A., 2005, Statistics in Medicine, 26, 3057
- Heinrich, J., and Lyons, L., 2007, Ann. Rev. Nucl. PartSci., 57, 145
- Heydorn, K., and Anglov, T., 2002, Accred. Qual. Assur., 7, 153
- Jarosik, N., et al., 2011, *ApJS*, 192, 14
- Jolliffe, I., 2002, *Principal Component Analysis*, 2<sup>nd</sup> ed., Springer, NY
- Kashyap, V.L., et al., 2008, Proc. SPIE, 7016, 7016P.1
- Kim, A.G., and Miquel, R., 2006, Astroparticle Physics, 24, 45
- Li, K.-H., et al., 1991, Statistica Sinica, 1, 65
- LIGO Collaboration 2010, *Nuclear Instruments and Methods in Physics Research A*, 624, 223
- Mandel, K.S., Wood-Vasey, W.M., Friedman, A.S., and Kirshner, R.P., 2009, *ApJ*, 704, 629
- Maness, et al., 2011, *ApJ*, 707, 1098

- Marshall, H., 2006, IACHEC, Lake Arrowhead, CA
- Mather, J.C., Fixsen, D.J., Shafer, R.A., Mosier, C., and Wilkinson, D.T., 1999, ApJ, 512, 511
- Meng, X.-L., 1994, Statistical Science, 9, 538
- Mohr, P.J., Taylor, B.N., and Newell, D.B., 2008, J. Phys. Chem. Ref. Data, 37, 3, 1187
- Osbourne 1991, International Statistical Review, 59, 3, 309
- Park, T., et al., 2008, ApJ, 688, 807
- Ramsay, J., and Silverman, B.W., 2005, *Functional Data Analysis*, Springer, 2<sup>nd</sup> ed., NY
- Refsdal, B. et al. 2009, Proc. of the 8th Python in Science Conference, (SciPy 2009), G. Varoquaux, S. van der Walt, J. Millman (Eds.), pp. 51-57 (2009)
- Rosset, C., et al. 2010, A&A, 520, 13
- Rubin, D.B. 1987, *Multiple Imputation for Nonresponse in Surveys*, J.Wiley & Sons, NY
- Rutherford, E.S., and Chadwick, J., 1911, Proc. Phys. Soc. London, 24, 141
- Schafer, J. L., 1997, *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York
- Schmelz, J.T. et al. 2009, ApJ, 704, 863
- Siemiginowska, A., et al., 2008, ApJ, 684, 811
- Simpson, G., and Mayer-Hasselwander, H., 1986, A&A, 162, 340
- Sundberg, Rolf, 1999, Scandinavian Journal of Statistics, 26, 161
- Taris, et al., 2011, A&A 526, A25
- Thoms, Maraston, & Johansson, 2010, Accepted for publication in MNRAS
- van Dyk, D., et al., 2001, ApJ, 548, 224

VIRGO Collaboration 2011, Classical and Quantum Gravity, 28, 2, 5005

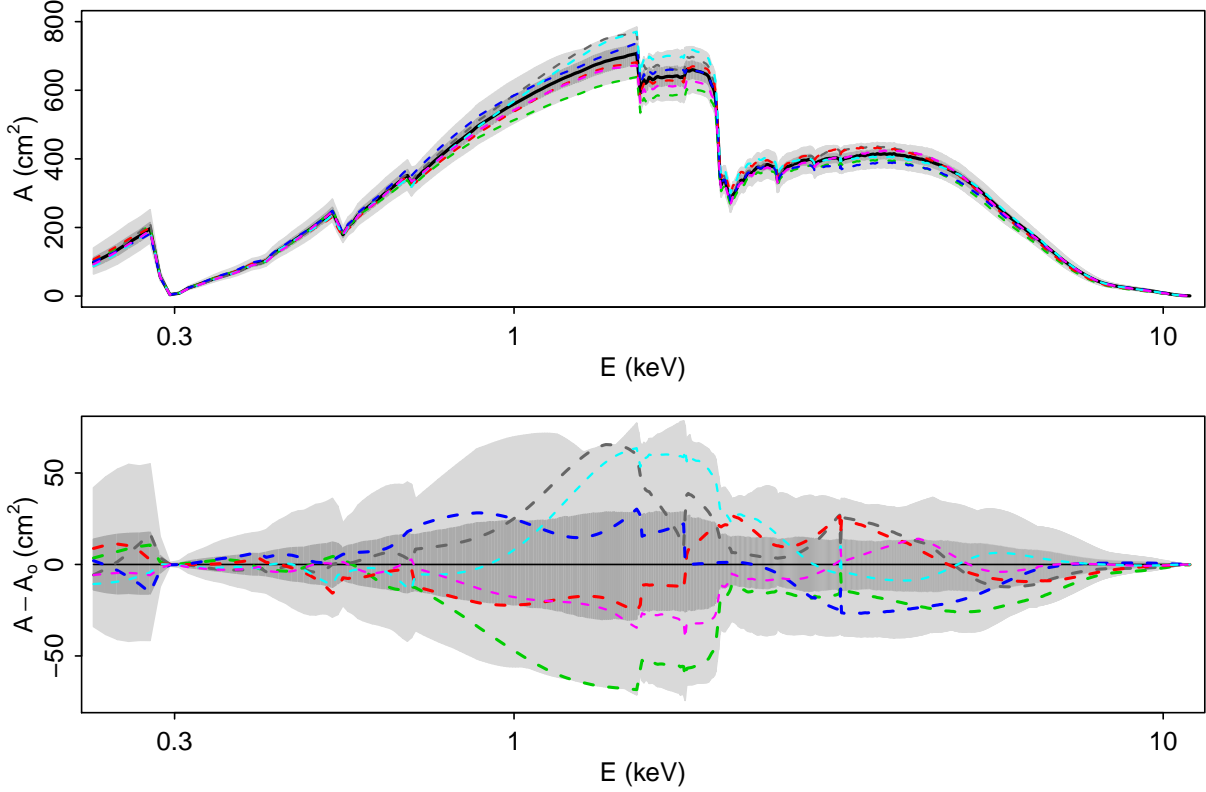


Fig. 1.— Uncertainty in ACIS-S effective area. In the upper panel the light gray area covers all 1000 effective area curves in the calibration sample of Drake et al. (2006) and the darker gray area covers the middle 68% of the curves in each energy bin. In addition six randomly selected curves are plotted as colored dashed curves and  $A_0$  is plotted as a solid black curve. The bottom panel is constructed in the same manner, but using  $A_l - A_0$ , in order to magnify the structure in  $\mathcal{A}$ . The curves in  $\mathcal{A}$  form a complex tangle that appears to defy any systematic pattern. As we shall see, we can use principle component analysis to form a concise summary of  $\mathcal{A}$ .



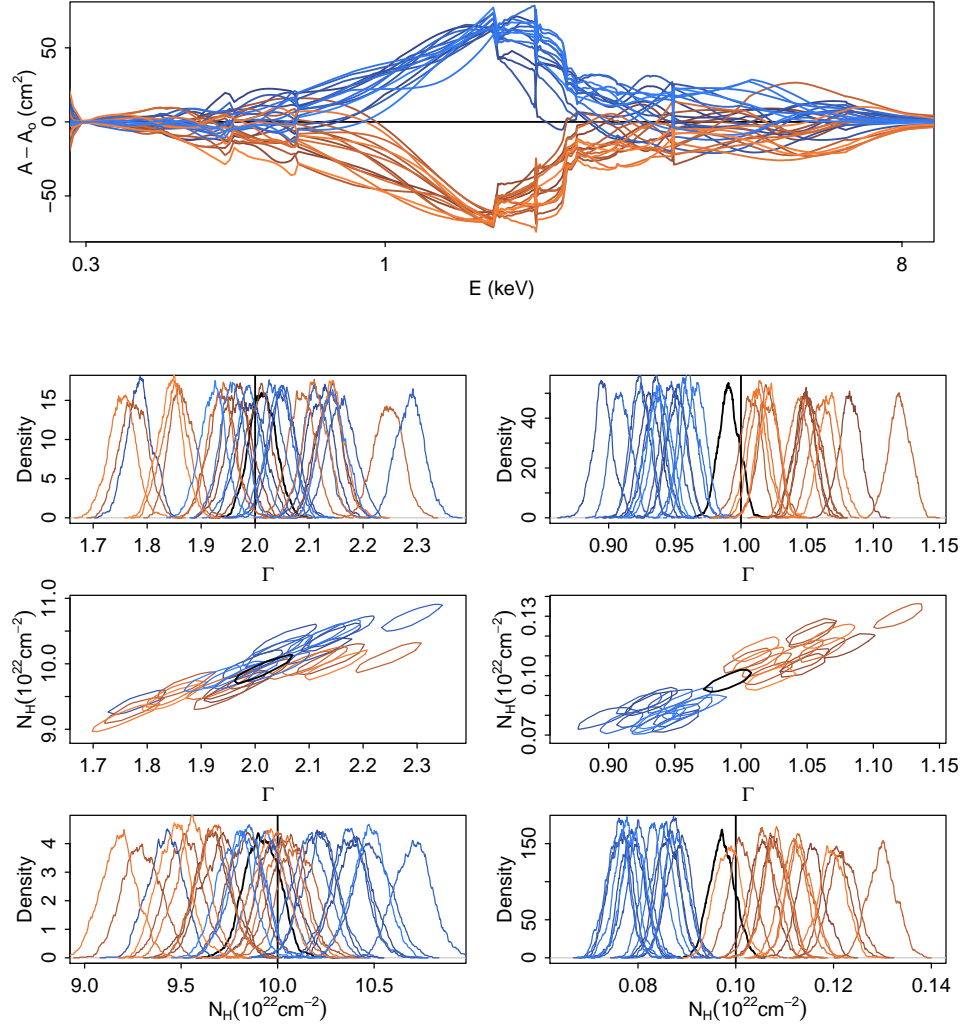


Fig. 2.— The Effect of Calibration Uncertainty on Fitted Parameters and Error Bars. The first panel plots the 15 effective area curves in  $\mathcal{A}$  with the largest maximum in blue and the 15 curves with the smallest maximum in red, each with  $A_0$  subtracted off. The solid black horizontal line at zero represents  $A_0$ . The two columns in the six lower panels correspond to SIMULATIONS 1 and 2, respectively and plot the posterior distributions of  $\Gamma$  and  $N_{\text{H}}$  using each of the 31 effective area curves in the first panel. The rows of the bottom six panels correspond to the posterior distribution of  $\Gamma$ , the 95.4% contour of the joint posterior distribution, and the posterior distribution of  $N_{\text{H}}$ . The colors of the plotted posterior distributions indicate the effective area curve that was used to generate the distribution. The solid vertical black lines in the the second and fourth rows indicate the values of the parameters used with  $A_0$  to generate SIMULATIONS 1 and 2. The effect of the choice of effective area curves on the posterior distributions is striking.

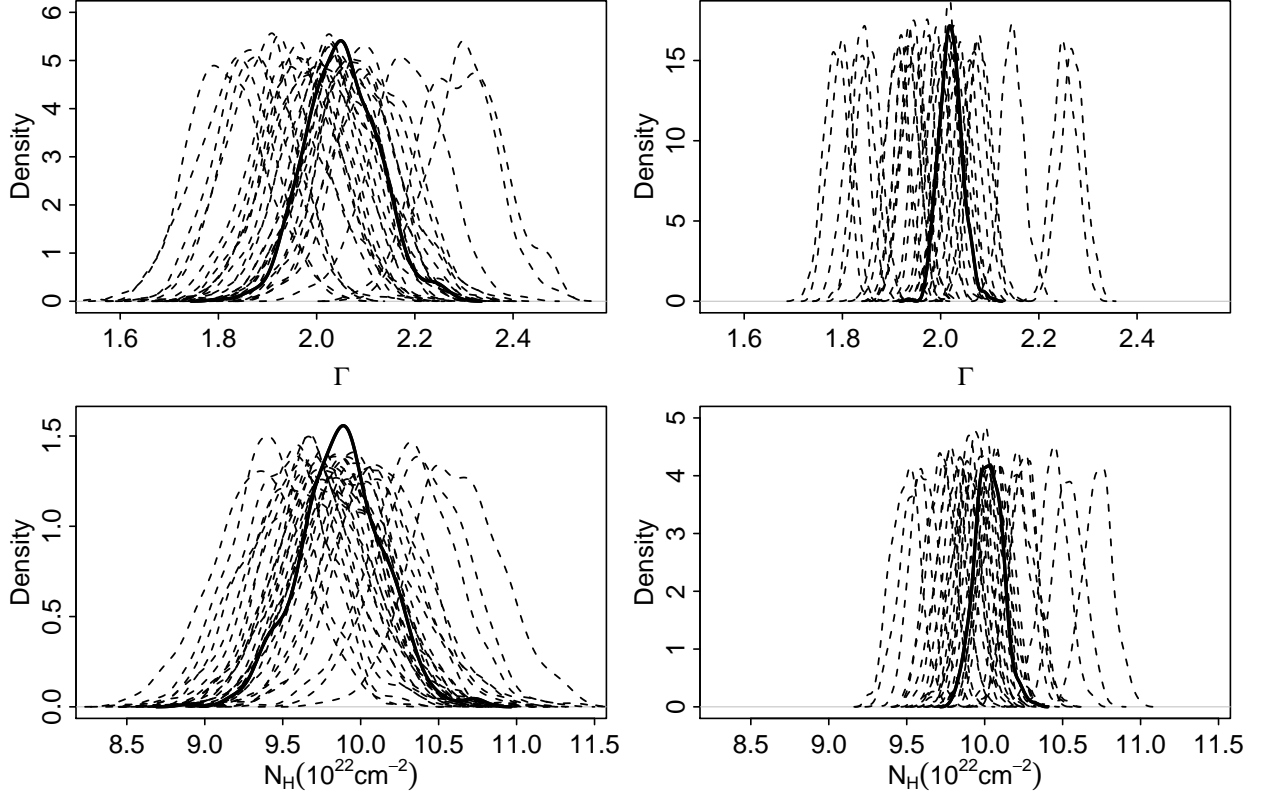


Fig. 3.— The Interaction Between Total Counts and Calibration Uncertainty. The four panels plot the marginal posterior distributions of  $\Gamma$  (row 1) and  $N_{\text{H}}$  (row 2) when fitting SIMULATION 3 (column 1 with  $10^4$  counts) and SIMULATION 1 (column 2 with  $10^5$  counts). The replicates in each panel correspond to 30 effective area curves randomly selected from  $\mathcal{A}$ . The posterior distributions plotted with solid lines were constructed using  $A_0$ . The statistical errors are smaller with the larger data set so that calibration errors are relatively more important.

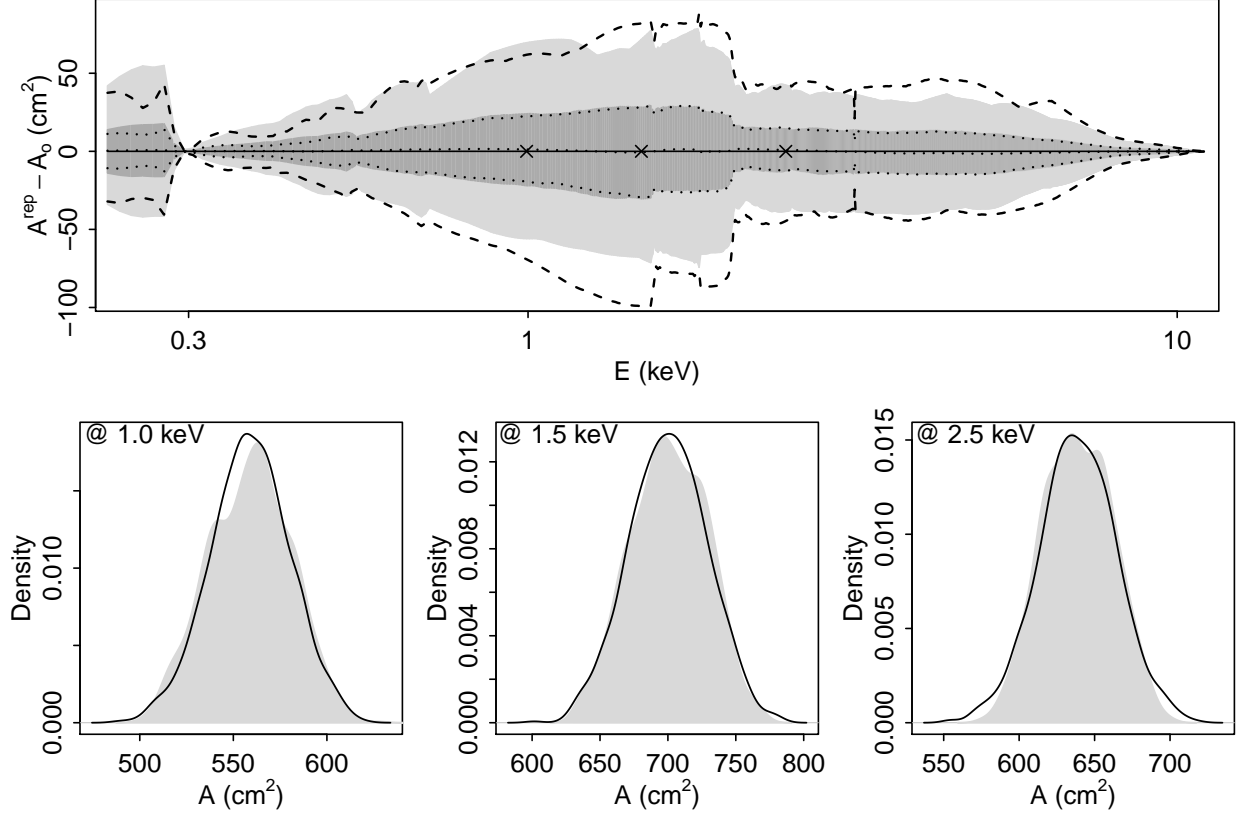


Fig. 4.— Summarizing the Calibration Sample Using PCA. The grey regions in the upper left panel are identical to those in the second panel of Figure 1 and give intervals for each energy bin that contain 100% and 68.3% of the calibration sample. The dashed and dotted lines outline intervals for each energy bin containing 100% and 68.3% of 1000 PCA replicates of the effective area, sampled using Equation 12. The correspondence between the calibration sample and the PCA sample is quite good, especially for the 68.3% intervals. The solid horizontal line is  $A_0$  and dotted line near it is the almost identical  $\bar{A}$ . The other three panels give histograms of the calibration sample (grey) and the PCA sample (solid line) in each of three energy bins, represented by  $\times$  signs in the first panel.

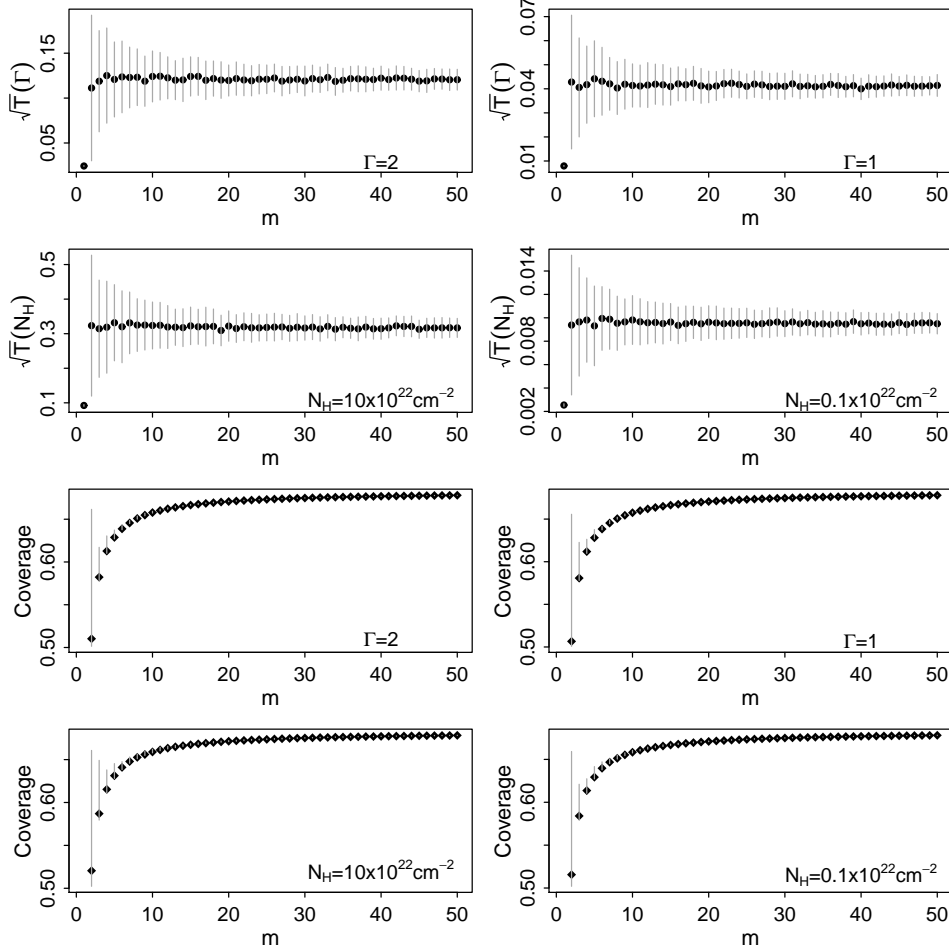


Fig. 5.— The Sensitivity of the Error Estimates on the Number of Imputations,  $M$ . We show the result of varying  $M$  on fits carried out for spectra from SIMULATION 1 (left column) and SIMULATION 2 (right column). For each  $M = m$ , we generate  $m$  effective area curves  $\{A_i^{\text{rep}}, i = 1, \dots, m\}$  using Equation 13, and carry out separate fits for each using *Sherpa*, and combine the the results of the fits using the multiple imputation combining rules (Equations 4–7). This gives us one value for the combined (statistical and systematic) error bar. We repeat this process 200 times for each  $m$  to investigate the variability of the computed error bar. The average computed errors (filled symbols) are shown for the power-law index  $\Gamma$  (top row) and the absorption column density  $N_{\text{H}}$  (second row) as a function of  $m$  along with the uncertainty on the errors due to sampling (thin vertical bars). The total error is grossly underestimated for  $m = 1$  (computed for only the default effective area), and the uncertainty on the error decreases for  $m > 1$ . Typically,  $M \approx 20$  is sufficient to obtain a reasonably accurate estimate of the total error. We also show the coverage fraction for the derived error bars for  $\Gamma$  (third row from the top) and  $N_{\text{H}}$  (bottom row). The coverage is small for small  $m$  because the degrees of freedom is small (see Equation 8) but asymptotically approaches Gaussian coverage of 0.683 for large  $m$ .

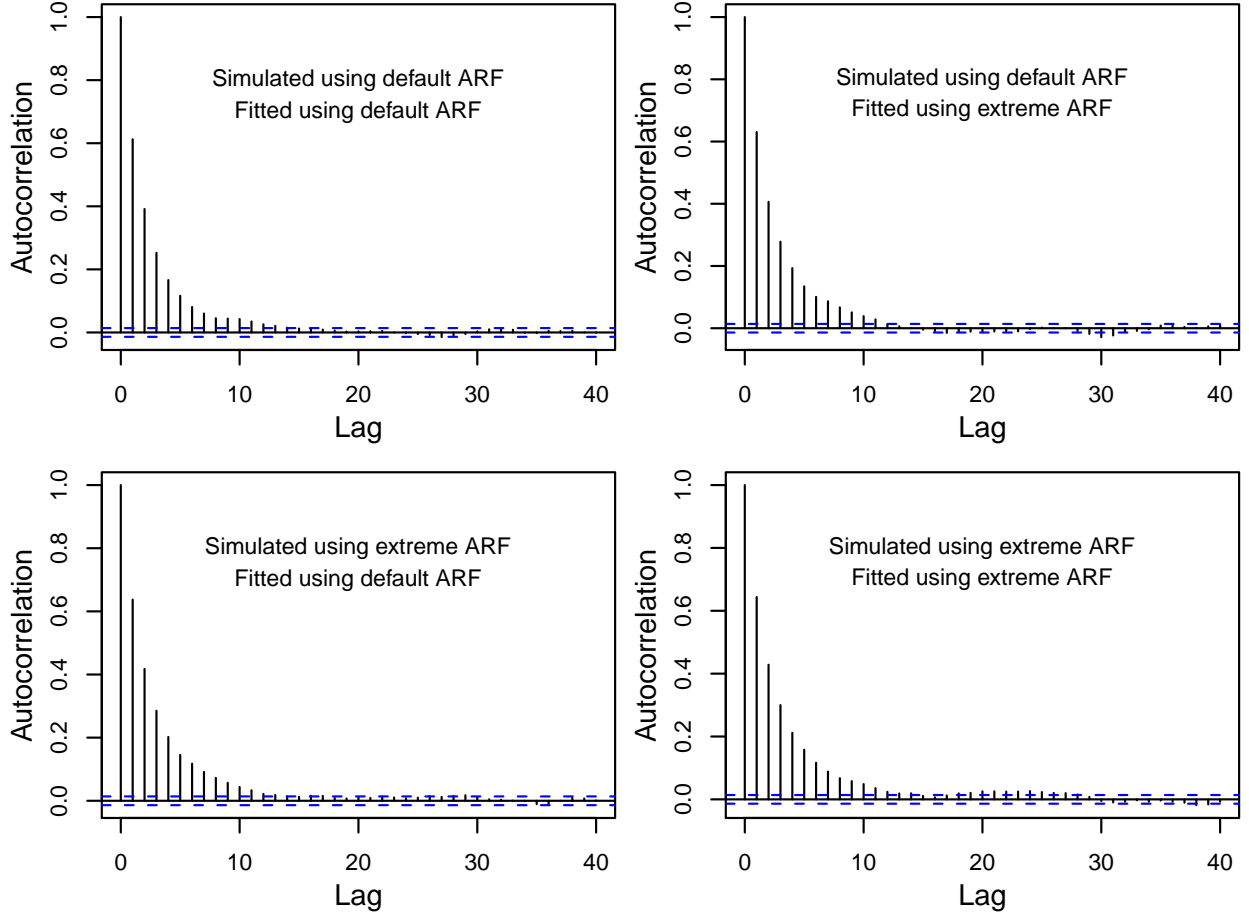


Fig. 6.— The Autocorrelation Function (ACF) of the Parameter Trace in MCMC Runs. The ACF for the spectral index  $\Gamma$  is shown for four cases, where a spectrum is simulated using one effective area curve and the fit is possibly carried out with another. This explores the dependence of the fitting methodology (codified in the routine `pyBLcXS`) on misspecified calibration. The top row shows the ACF for SIMULATION 1 (generated using “default” effective area curve; see Table 2) and the bottom row for SIMULATION 5 (generated using an “extreme” effective area curve). The diagonal plots show the ACF when the “correct” effective curve is used to fit the spectrum, i.e., the same curve as was used to generate it, and the cross-diagonal plots show the case when the fitting is carried out using a different effective area curve. The cases in the left column both use the “default” effective area to fit the simulated spectra, and the cases in the right column both use the “extreme” curve. The autocorrelation functions demonstrate that  $\Gamma^{(k)}$  and  $\Gamma^{(k+10)}$  are essentially uncorrelated regardless of whether the correct effective area curve was used in the fit or not. Thus, we set  $I = 10$  in our pragmatic Bayesian samplers.

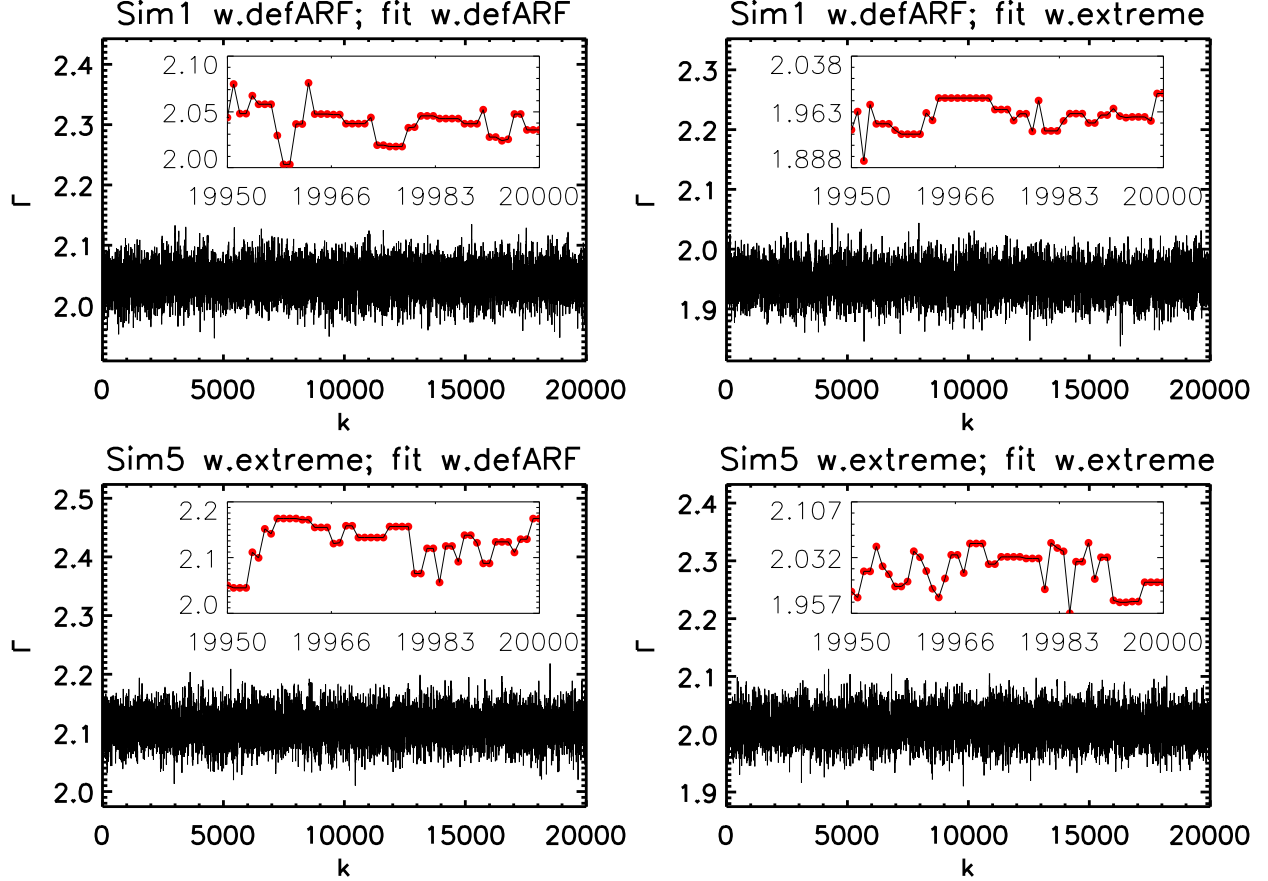


Fig. 6.— (CONTD.) The parameter traces for the spectral index  $\Gamma$ , shown for same cases as the autocorrelation cases shown before. While the autocorrelation determines the “stickiness” of the MCMC iterations, the time series demonstrates that choosing misspecified calibration files does not have any effect on the convergence of the solutions. The traces are shown in the same order as before, for all iterations  $k$ . The inset shows the last 50 iterations, with  $\Gamma^{(k)}$  denoted by filled circles, and consecutive iterations connected by thin straight lines. The necessity of using  $I \gg 1$  is apparent in the slow changes in the values of  $\Gamma^{(k)}$ .

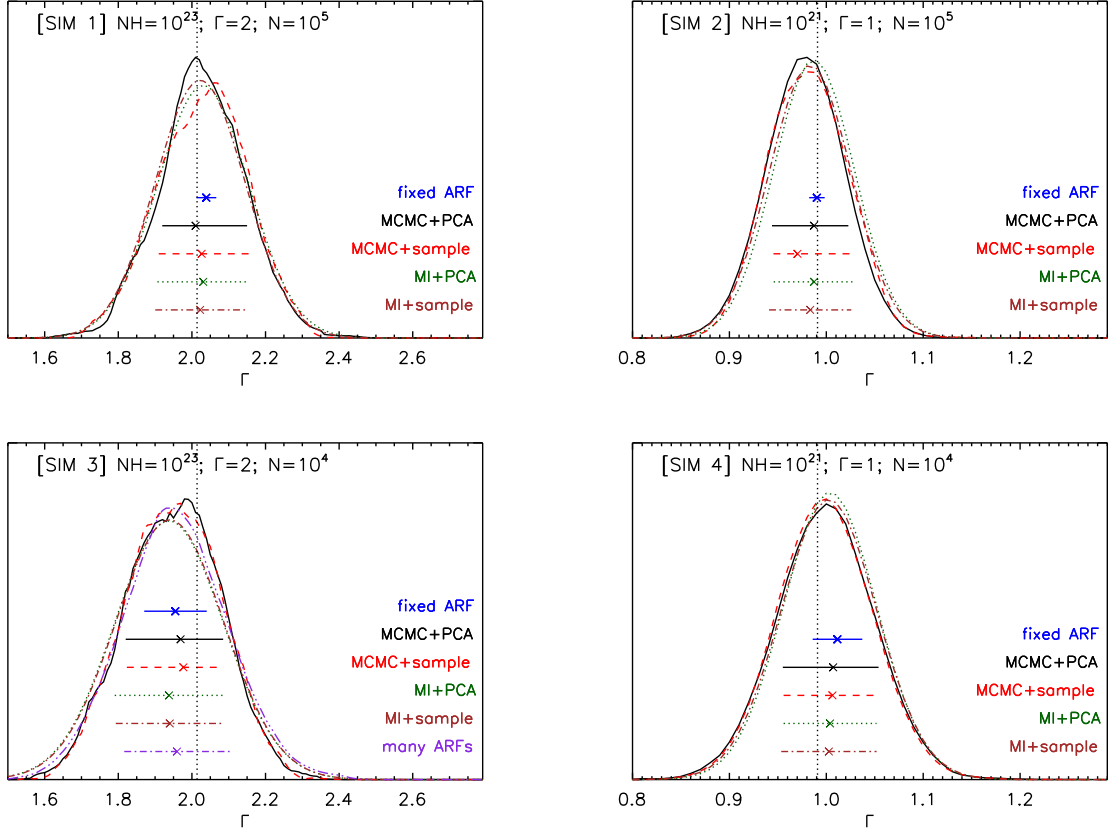


Fig. 7.— Comparing the Algorithms in §5 as Applied to the Simulated Spectra 1-4 in Table 2. These are spectra which are generated using the default effective area. The “true” value of the power-law index parameter that was used to generate the simulated spectra is shown as the vertical dashed line. For each simulation, posterior probability density functions of the power-law index parameter are computed using the pragmatic Bayesian with PCA (black solid curve; §4.2.4), pragmatic Bayesian with sampling from  $\mathcal{A}$  (red dashed curve; §4.2.3), Multiple Imputation with PCA (green dotted curve; §4.1.2), Multiple Imputation with samples from  $\mathcal{A}$  (brown dot-dashed curve; §4.1.1), and the combined posteriors from individual runs using the full sample  $\mathcal{A}$  (purple dash-dotted curve). Results for the column density parameter  $N_{\text{H}}$  are similar. We use  $M = 20$  samples for multiple imputation. The density curves are obtained from smoothed histograms of MCMC traces from `pyBLoCXs` for the Bayesian cases, and are Gaussians with the appropriate mean and variance obtained via fitting with `XSPEC v12` for the Multiple Imputation cases. Also shown are the 68% equal-tail intervals as horizontal bars, with the most probable value of the photon index indicated with an ‘x’ for each of these case, and additionally for the case where a fixed effective area was used to obtain only the statistical error. Note that in all cases, fitting with the default effective area alone leads to an underestimate of the true uncertainty in the fitted parameter.

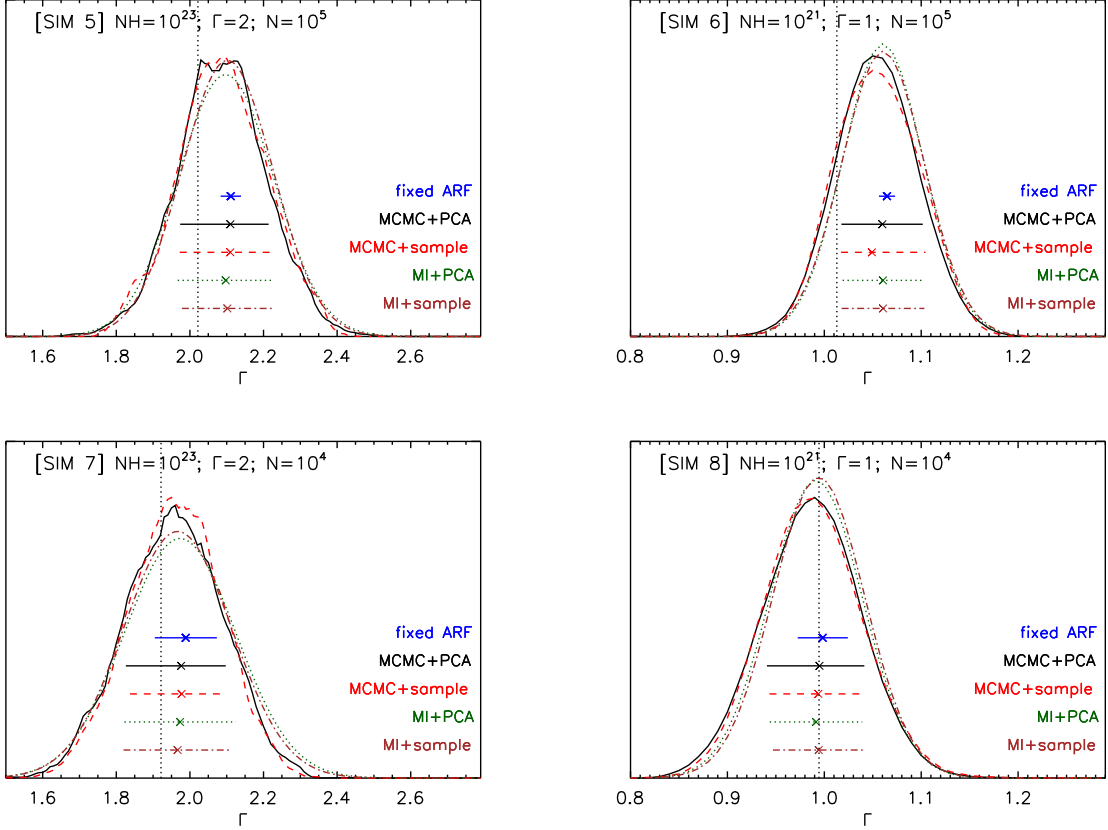


Fig. 7.— (CONTD.) For Simulated Spectra 5-8 in Table 2. These are spectra which are generated using an extreme instance of an effective area from  $\mathcal{A}$ . The fits when only one effective area is used are done with the default effective area. Note that in many cases, not incorporating the calibration uncertainties results in intervals for the parameter which does not contain the true value.



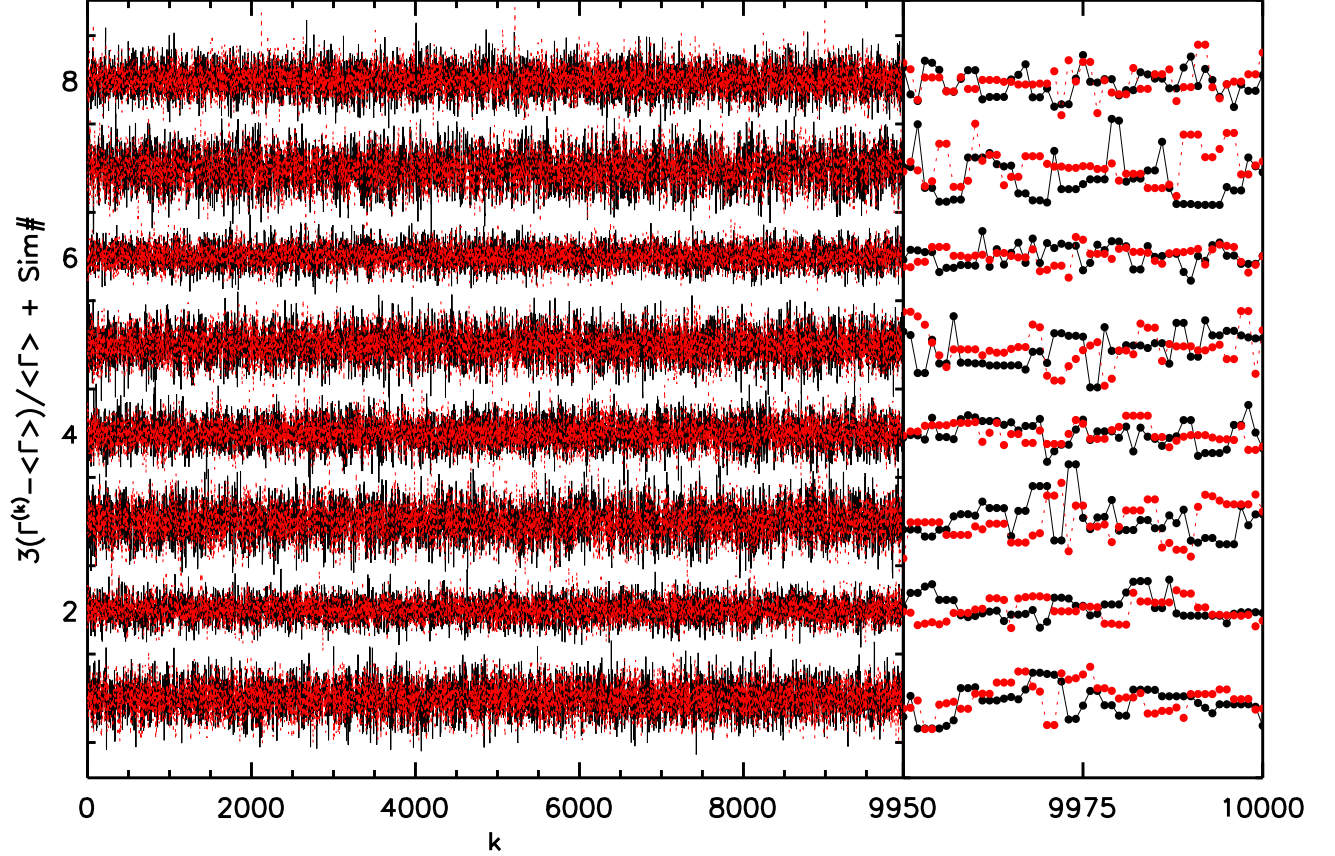


Fig. 7.— (CONTD.) Parameter traces for the spectral index  $\Gamma$  for each of the 8 simulations. All the simulations are shown on the same plot, rescaled (to depict the fractional deviation from the mean, inflated by a factor of 3) and offset (by an integer corresponding to the number assigned to the simulation) for clarity. The traces for both the MCMC+PCA (pragmatic Bayesian algorithm using PCA to generate new effective areas; solid black lines) and MCMC+sample (pragmatic Bayesian algorithm with sampling from  $\mathcal{A}$ ; dotted red lines) are shown, with the latter overlaid on the former. The last 50 iterations are shown zoomed out in the abscissa for clarity, and shows each transformed  $\Gamma^{(k)}$  as filled circles, connected by thin lines of the corresponding style and color. Note that all iterations  $k$  are shown, but in the calculations of the posterior probability distributions, only every  $I^{th}$  iteration, where  $I = 10$ , is used (see Figure 6).

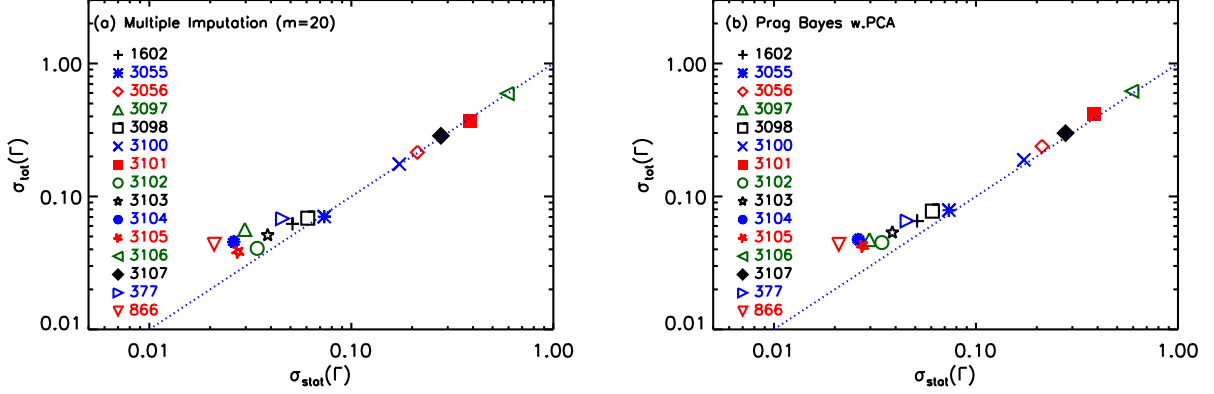


Fig. 8.— Comparison of the Statistical Error with the Total Error Including Effective Area Uncertainties for Different Methods of Evaluating Them. Results of fits to a sample of 15 radio loud quasars (Siemiginowska et al. 2008; see §5.4) are shown. The abscissae represent the statistical uncertainty  $\sigma_{\text{stat}}$  as derived by adopting a fixed, nominal effective area, and fit with absorbed power-law models using *CIAO*/Sherpa (stronger sources tend to have smaller error bars). They are compared with the total error,  $\sigma_{\text{tot}}$  derived using (a) the Multiple Imputation combining rule (§4.1.2) with *CIAO*/Sherpa ( $M = 20$ ), and (b) the pragmatic Bayesian method with PCA (§4.2.4), with *pyBLcXS*. (Similar results are obtained when using the pragmatic Bayesian method for the full sample of effective areas.) The different symbols correspond to the analysis carried out for different observations. The dotted line represents equality, where the total error is identical to the statistical error. The systematic error cannot be ignored when the statistical error is small, and represents the limiting accuracy of a measurement.

Table 1. Glossary of symbols used in the text

Symbol	Description
$A$	effective area (ARF) curve
$A^{\text{rep}}$	replicate $A$ generated from PCA representation of the calibration sample
$A_0$	the default effective area curve.
$A_0^*$	the observation specific effective area curve.
$A_l$	effective area curve $l$ in the calibration sample
$\mathcal{A}$	a set of effective areas, the calibration sample
$\delta\bar{A}$	average offset of $\mathcal{A}$ from $A_0$
$B$	the between imputation (or systematic) variance of $\hat{\theta}$ .
$B_{mm}$	diagonal element $m$ of $B$
$E$	energy of incident photon
$E^*$	energy channel at which the detector registers the incident photon
$e_j$	random variate generated from the standard Normal distribution
$f_l$	fractional variance of component $l$ in the PCA representation
$I$	number of inner iterations in pyBLoCXS, typically 10
$J$	number of components used in PCA analysis, here 17
$j$	principal component number or index
$(k)$	the superscript indicates the running index of random draws
$\mathcal{K}$	an MCMC kernel
$\mathcal{K}_{\text{pyB}}$	the MCMC kernel used in PyBLoCKS
$L$	number of replicate effective area curves in calibration sample
$l$	replicate effective area number or index, or principal component number

Table 1—Continued

Symbol	Description
$m$	imputation number or index
$M$	number of imputations
$\mathcal{M}$	response of a detector to incident photons, see Equation 1
$p$	objective function (posterior distribuiton, likelihood, or perhaps $\chi^2$ )
$P$	point spread function (PSF)
$R$	energy redistribution matrix (RMF)
$r_l^2$	eigenvalue or PC coefficient of component $l$ in the PCA representation
$S$	astrophysical source model
$T$	total variance of $\hat{\theta}$ .
$v_l$	eigen- or feature-vector for component $l$ in the PCA representation
$W$	the within imputation (or statistical) variance of $\hat{\theta}$ .
$W_{mm}$	diagonal elements $m$ of $W$
$\mathbf{x}$	true sky location of photons
$\mathbf{x}^*$	locations of incident photons as registered by detector
$Y$	data, typically used here as counts spectra in detector PI bins
$Z$	data and physical calculations used by calibration scientists
$\theta$	model parameter of interest
$\hat{\theta}$	estimate of $\theta$
$\hat{\theta}_m$	estimate of $\theta$ corresponding to imputed effective area $m$
$\text{Var}(\hat{\theta}_m)$	estimates variance of $\hat{\theta}_m$
$\sigma_{\text{stat}}$	$\sqrt{W}$ , representing the statistical error on $\theta$

Table 1—Continued

Symbol	Description
$\sigma_{\text{tot}}$	$\sqrt{T}$ , representing the total error on $\theta$
$\xi$	a sum of the smaller components, J+1 to L in the PCA representation

Table 2: The Eight Simulations Used to Compare the Four Algorithms Described in §4.

	Effective Area		Nominal Counts		Spectral Model	
	Default	Extreme	$10^5$	$10^4$	Hard <sup>†</sup>	Soft <sup>‡</sup>
SIMULATION 1	X		X		X	
SIMULATION 2	X		X			X
SIMULATION 3	X			X	X	
SIMULATION 4	X			X		X
SIMULATION 5		X	X		X	
SIMULATION 6		X	X			X
SIMULATION 7		X		X	X	
SIMULATION 8		X		X		X

<sup>†</sup>An absorbed powerlaw with  $\Gamma = 2$ ,  $N_{\text{H}} = 10^{23}/\text{cm}^2$

<sup>‡</sup>An absorbed powerlaw with  $\Gamma = 1$ ,  $N_{\text{H}} = 10^{21}/\text{cm}^2$